



University
of Glasgow

Pieszko, Marta (2015) *Molecular regulation of the macroschizont to merozoite differentiation in Theileria annulata*. PhD thesis.

<http://theses.gla.ac.uk/6079/>

Copyright and moral rights for this thesis are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the Author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the Author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

**Molecular regulation of the macroschizont to
merozoite differentiation
in *Theileria annulata***

MARTA PIESZKO, M.Sc.



University
of Glasgow

**Submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy**

**College of Medical, Veterinary and Life Sciences
School of Veterinary Medicine
Institute of Biodiversity, Animal Health and Comparative Medicine
University of Glasgow
February, 2015**

© Marta Pieszko, 2015

Author's Declaration

I hereby declare that the work presented in this thesis is entirely my own original work and carried out with the help of those people mentioned in acknowledgements. Where other sources of information have been used, they have been acknowledged. The research work presented in this thesis has not been previously submitted to any other university for the award of a degree.

Marta Pieszko

September, 2014

Abstract

Theileria annulata is an intracellular, tick-transmitted apicomplexan parasite, which causes tropical theileriosis in cattle. It undergoes a complex life cycle with several distinct stages occurring within the bovine host and tick vector. ApiAP2 proteins are key candidate transcription factors for regulation of stage-specific gene expression across apicomplexans. They are differentially expressed in specific developmental stages and certain ApiAP2s bind specifically to unique DNA sequence motifs. Identification of stage-specific expression of putative transcriptional regulators, the motifs they bind to and potential target genes provided the rationale for this study to understand the molecular mechanisms that control stage differentiation to the merozoite in *T. annulata*.

The results demonstrated that *T. annulata* ApiAP2s show marked differences in expression levels during the parasite life cycle. ApiAP2 target DNA motifs orthologous to those in *Plasmodium* and *Cryptosporidium* were also discovered in *Theileria* intergenic regions, indicating that the genes downstream are potential targets of *Theileria* ApiAP2s. These motifs were also found in upstream regions of up-regulated TaApiAP2 genes, suggesting possible auto-regulation and an interaction network of ApiAP2 transcription factors. Importantly ApiAP2 fusion proteins up-regulated during differentiation to the merozoite stage bound to their predicted specific DNA motifs validating that ApiAP2 DNA-binding domain structure is conserved across Apicomplexa genera.

Evidence was also produced that AP2 proteins play important roles in steps that commit a cell to differentiate: TA13515D is the orthologue of the AP2G factor in *Plasmodium* that is a major regulator of gametocytogenesis: TA16485 may be involved in down-regulation of genes during merogony and expression of TA11145 at a higher level in a cell line competent for merogony relative to a line severely attenuated indicated involvement in regulation of this differentiation step. Discovery of multiple nuclear factors binding to a 2x(A)CACAC(A) motif implicated in autoregulation of TA11145, together with phylogenetic evidence for a clade of related domains that bind this motif suggest that multiple competing ApiAP2s may operate to regulate stochastic commitment to merozoite

production. Based on this data an updated stage differentiation model has been generated, with up regulation of the TA11145 gene a key event.

A C-box motif association with genes implicated in establishment of the transformed host cell (TashAT, SVSP) suggests it could be important for deregulation of this event as the parasite undergoes stage differentiation. In contrast the inverse G-box was found associated with genes up-regulated from merozoite to piroplasm. EMSA analysis of parasite nuclear extract with a G/C-box motif probe showed that the motif is an active binding site for a stage regulated nuclear factor. Specific binding of candidate TA12015 protein to the G/C-box motif was unable to be confirmed.

Taken together, these results provided evidence that ApiAP2 proteins are regulators of stage-specific gene expression in *T. annulata*. They also provide insight into probable ApiAP2 interaction networks and support the postulation of a differentiation mechanism conserved across the Apicomplexa. Finally, the data suggests that this mechanism is stochastic and is likely to occur via a positive feedback loop generating a threshold that commits the cell to differentiate to the next stage of the life cycle.

Table of contents

Author`s declaration.....	2
Abstract.....	3
Table of contents.....	5
List of figures	12
List of tables	18
Acknowledgements.....	20
List of symbols and abbreviations.....	22

CHAPTER 1

1	INTRODUCTION.....	28
1.1	Introduction	28
1.2	<i>Theileria</i> and closely-related Apicomplexa species	29
1.3	Economic importance and pathogenesis of tropical theileriosis	32
1.4	<i>Theileria annulata</i> control strategies	34
1.5	Life-cycle of <i>T. annulata</i>	36
1.5.1	Bovine stages of the <i>T. annulata</i> life-cycle.....	37
1.5.2	Tick stages of the <i>T. annulata</i> life-cycle.....	41
1.6	A stochastic model of stage differentiation in <i>T. annulata</i>	41
1.7	Regulation of gene expression in apicomplexan parasites	44
1.8	Aims and objectives.....	48

CHAPTER 2

2	INVESTIGATION OF <i>THEILERIA ANNULATA</i> GENE EXPRESSION DURING STAGE DIFFERENTIATION FROM MACROSCHIZONT TO MEROZOITE	52
2.1	INTRODUCTION.....	52
2.2	MATERIAL AND METHODS.....	59
2.2.1	Microarray design	59
2.2.1.1	Statistical analysis of microarray data	60
2.2.2	Rank Product analysis of differential gene expression	61
2.2.3	Cell lines and <i>in vitro</i> culture of <i>T. annulata</i>	61
2.2.3.1	Cryopreservation of cell lines	62
2.2.3.2	<i>T. annulata</i> differentiation time-course.....	62

2.2.4	Total RNA isolation and processing.....	63
2.2.5	Reverse-Transcription Polymerase Chain Reaction	64
2.2.5.1	Primer design and analysis	65
2.2.5.2	Semi-quantitative RT-PCR reagents and cycling conditions	66
2.2.5.3	Gel electrophoresis and quantification of RT-PCR products.....	67
2.2.6	Two-step quantitative Reverse-Transcription PCR.....	67
2.2.6.1	cDNA synthesis	68
2.2.6.2	Housekeeping genes	69
2.2.6.3	qRT-PCR reaction parameters.....	69
2.2.6.4	qRT-PCR data analysis.....	69
2.3	Results.....	71
2.3.1	Hierarchical clustering of <i>T. annulata</i> genes	71
2.3.2	Scatter plot analysis	73
2.3.3	Pair-wise comparison of genes differentially expressed between <i>T. annulata</i> life cycle stages	76
2.3.3.1	Identification of genes differentially expressed between the macroschizont to merozoite stage	77
2.3.3.2	Identification of genes differentially expressed between the sporozoite to macroschizont stage	81
2.3.3.3	Identification of genes differentially expressed between the merozoite to piroplasm stage.....	81
2.3.3.4	Identification of genes differentially expressed between the piroplasm to sporozoite stage	82
2.3.4	Microarray life cycle expression profiles of putative transcription factors of <i>T. annulata</i>	83
2.3.5	Differentiation from macroschizont to merozoite stage of <i>T. annulata</i> in vitro.....	89
2.3.6	Validation of differentially regulated genes by semi-quantitative RT-PCR	91
2.3.7	Validation of differential expression patterns of selected potential transcription factors by RT-PCR	94
2.3.8	Validation of expression pattern of selected up-regulated and down-regulated genes by qRT-PCR	96
2.4	Discussion	101

CHAPTER 3

3	IDENTIFICATION AND ANALYSIS OF <i>T. ANNULATA</i> CIS-ACTING PROMOTER ELEMENTS.....	110
3.1	INTRODUCTION.....	110
3.2	MATERIAL AND METHODS	119
3.2	Motif discovery in upstream regions of <i>T. annulata</i> co-expressed genes and gene families	119
3.2.1	Cross-species MEME analysis of the TashAT cluster	117
3.2.2	Analysis of the intergenic region of TA03110 (Tash-a) across <i>T. annulata</i> , <i>T. parva</i> and <i>T. orientalis</i>	121
3.2.3	Statistical enrichment analysis of 4C-box and 5C-box motifs within <i>T. annulata</i> intergenic regions	121
3.2.4	ApiAP2 domain conservation analysis across <i>Theileria</i> and <i>Plasmodium</i> species	122
3.2.5	Distribution and enrichment analysis of <i>Plasmodium</i> motifs identified in <i>Theileria</i> genome	123
3.2.6	Motif enrichment analysis of the upstream regions of ApiAP2 genes in <i>Theileria</i>	123
3.2.7	Temporal expression pattern analysis of ApiAP2 genes and their potential target genes.....	124
3.3	Results	125
3.3.1	DNA motifs found in upstream regions of genes co-expressed from the macroschizont to merozoite stage	125
3.3.2	Cross-species analysis for the TashAT cluster	133
3.3.2.1	MEME analysis of the intergenic region of TA03110 (Tash-a) across <i>T. annulata</i> , <i>T. parva</i> and <i>T.orientalis</i>	135
3.3.3	Statistical enrichment analysis of 4C-box motif	137
3.3.4	Enrichment analysis of 5C-box motif.....	141
3.3.5	ApiAP2 protein conservation across <i>Theileria</i> and <i>Plasmodium</i> species .	144
3.3.6	Motif distribution in the genome of <i>T. annulata</i> and enrichment analysis of motifs bound by <i>P. falciparum</i> ApiAP2s.....	149
3.3.6.1	G-box/C-box motif analysis	152
3.3.6.2	GTGTAC, TCTACA and ACACAC motif analysis.....	155
3.3.7	Motifs found in upstream regions of ApiAP2 genes of <i>Theileria</i>	159

3.3.8 Temporal expression pattern analysis of ApiAP2 genes and their potential target genes	160
3.4 Discussion	164

CHAPTER 4

4 APIAP2 GENES AS CANDIDATE TRANSCRIPTION FACTORS FOR <i>T. ANNULATA</i>	174
4.1 INTRODUCTION	174
4.2 Material and Methods	181
4.2.1 ApiAP2 proteins: general characteristics and selected domain conservation analysis across <i>Theileria</i> and <i>Plasmodium</i> species	181
4.2.2 ApiAP2 domain structure prediction	181
4.2.3 Generation of selected <i>Theileria</i> ApiAP2 GST-fusion proteins.....	182
4.2.3.1 PCR amplification and cloning of ApiAP2 domains from <i>T. annulata</i> genomic DNA.....	183
4.2.3.2 Construction of recombinant plasmids	184
4.2.3.3 Bacterial transformation	184
4.2.3.4 Test induction of GST fusion proteins	185
4.2.3.5 Enhancing solubility of expressed recombinant proteins in <i>E. coli</i>	186
4.2.3.6 Protein expression for large-scale production	186
4.2.3.7 GST- fusion protein purification	187
4.2.4 Chemiluminescent Electrophoretic Mobility Shift Assay (EMSA).....	188
4.2.4.1 EMSA with Polyamide inhibitors	191
4.2.5 Comparison of expression data from the D7 cell line and the D7B12 cell line attenuated for merogony	192
4.2.6 Immunofluorescent Antibody Test (IFAT)	192
4.3 Results	194
4.3.1 <i>T. annulata</i> ApiAP2 family, general characteristics	194
4.3.2 ApiAP2 proteins structure prediction.....	197
4.3.3 Validation of GST-TA13515D AP2 domain binding to its putative binding motif by EMSA.....	202
4.3.3.1 EMSA ASSAY OF GST-TA13515D WITH PA INHIBITORS	206
4.3.4 EMSA results of the GST-TA11145D fusion protein and its putative binding motif	208

4.3.5	EMSA results of GST-TA12015 and TA16485D fusion proteins and their putative binding motif	210
4.3.6	Detection by EMSA of down-regulated <i>T. annulata</i> parasite nuclear factors that bind to the C-box oligonucleotide motif probe	213
4.3.7	Multiple AP2 domains are predicted to bind (A)CACAC(A) type motifs.	215
4.3.7.1	Detection by EMSA of <i>T. annulata</i> parasite nuclear factors that bind to TA11145 oligonucleotide motif probes	218
4.3.7.2	Comparative analysis of two putative (A)CACAC(A)-binding ApiAP2 proteins in cell lines competent or attenuated for differentiation to the merozoite	223
4.3.8	IFAT results	227
4.4	Discussion	228

CHAPTER 5

5. GENERAL DISCUSSION	237
References	245

Appendices

Appendix 1: MATERIAL AND METHODS	273
1.1. List of buffers and solution commonly used in the experimental protocols of this thesis.....	273
1.2. Semi-quantitative RT-PCR reaction components and cycling conditions ...	276
1.3. cDNA synthesis reaction components and thermal cycling conditions ...	277
1.4. Table of primers used for semi-quantitative RT-PCR.	278
1.5. SYBR Green qRT-PCR reaction details and thermal cycling parameters ...	279
1.6. PCR conditions with Pfu polymerase and reaction components	279
1.7. Table of primers used for quantitative real time RT-PCR	280
1.8. Table of primers with restriction sites used for cloning.	280
1.9. Table of oligos used for EMSA experiment.	281
Appendix 2: Rank product results from Chapter 2	283
2.1. Full list of genes up-regulated from macroschizont to merozoite obtained by Rank Product (n=152, FDR≤0.05)	283
2.2. Full list of genes down-regulated from macroschizont to merozoite obtained by Rank Product (n=115, FDR≤0.05)	286

2.3. Full list of genes up-regulated from sporozoite to macroschizont obtained by Rank Product (n=67; FDR≤0.05)	291
2.4. Full list of genes down-regulated from sporozoite to macroschizont obtained by Rank Product (n=30; FDR≤0.05)	293
2.5. Full list of genes up-regulated from merozoite to piroplasm obtained by Rank Product (n=24; FDR≤0.05)	294
2.6 Full list of genes down-regulated from merozoite to piroplasm obtained by Rank Product (n=20; FDR≤0.05)	295
2.7. Full list of genes up-regulated from piroplasm to sporozoite obtained by Rank Product (n=57; FDR≤0.05)	295
2.8. Full list of genes down-regulated from piroplasm to sporozoite obtained by Rank Product (n=35; FDR≤0.05)	297
Appendix 3. <i>T. annulata</i> DNA motifs from Chapter 3	299
3.1 Top 3 motifs identified by MEME in the 5` intergenic regions of the 100 most up-regulated genes during differentiation of macroschizont to merozoite stage differentiation in <i>T. annulata</i>	299
3.2 Top 3 motifs identified by MEME in 5' intergenic regions of the 100 most down-regulated genes during differentiation of the macroschizont to merozoite stage in <i>T. annulata</i>	299
3.3. Top 3 motifs identified by MEME in the 3` intergenic regions of the 100 most up-regulated genes during differentiation of macroschizont to merozoite stage differentiation in <i>T. annulata</i>	300
3.4. Top 3 motifs identified by MEME in 3' intergenic regions of the 100 most down-regulated genes during differentiation of the macroschizont to merozoite stage in <i>T. annulata</i>	300
3.5. Top 3 motifs identified by MEME in 5' intergenic regions of all <i>T. annulata</i> ApiAP2 genes	301
3.6. Top 3 motifs identified by MEME in 5' intergenic regions of all other <i>T. annulata</i> transcription factors	301
3.7. Top 3 motifs identified by MEME in 5' intergenic regions of <i>T. annulata</i> genes constitutively expressed in macroschizont to merozoite stage	302
3.8. Full alignment of upstream sequences of <i>T. annulata</i> and <i>T. parva</i> orthologues of TA12015.	303
3.9. Full alignment of upstream sequences of <i>T. annulata</i> and <i>T. parva</i> orthologues of TA11145	304

3.10. Full alignment of upstream sequences of <i>T. annulata</i> and <i>T. parva</i> orthologues of TA13515	306
3.11. Full alignment of upstream sequences of <i>T. annulata</i> and <i>T. parva</i> orthologues of TA16485	307
Appendix 4 - qRT-PCR results of (A)CACAC(A) binding TaApiAP2s	308
4.1. Comparison of relative expression levels of TA11145 vs TA07100 in D7 and D7B12 cell lines	308

Figure 2.16. Validation of temporal profile of most down-regulated potential transcription factors by semi-quantitative RT-PCR	96
Figure 2.17. Temporal expression profile derived from microarray data of two heat shock proteins (TA10720 and TA11610) used as constitutive controls for qRT-PCR	96
Figure 2.18. QRT-PCR results of A. TA13515 and B.TA11145: using RNA representing macroschizont (Day 0) cultures, and a differentiation time course to merozoite production (Day 7 and 9) and the piroplasm stage	98
Figure 2.19. QRT-PCR results of A. TA10735 and B. TA15705 from macroschizont to merozoite stage	100
Figure 3.1. Top three MEME-derived motifs in <i>T. parva</i> and their most similar motifs in <i>T. annulata</i>	112
Figure 3.2. <i>P. falciparum</i> ApiAP2 domains and their corresponding target DNA motifs	115
Figure 3.3. UPGMA tree of the ApiAP2 domain in Apicomplexa	116
Figure 3.4. ATCCCCCAT motif found in the 5' intergenic regions of 100 most down-regulated genes during differentiation of the macroschizont to merozoite stage in <i>T. annulata</i>	127
Figure 3.5. Distribution of the ATCCCCCAT motif within intergenic regions of down-regulated genes	128
Figure 3.6. Motif TCCCCAT (4C-box) found in 5' IGR of TashAT family genes	129
Figure 3.7. TGTGT motif identified in the 5' intergenic regions of the 100 most up-regulated genes during macroschizont to merozoite stage differentiation in <i>T. annulata</i>	131
Figure 3.8. Distribution of the TGTGT motif within 5' intergenic regions of up-regulated genes	132
Figure 3.9. Cross species MEME analysis of the intergenic regions of <i>T. annulata</i> and <i>T. parva</i> TashAT cluster	134
Figure 3.10. Distribution of the 4C-box motif within intergenic regions of down-regulated genes	134
Figure 3.11. Cross species MEME analysis of the intergenic regions of <i>T. annulata</i> and <i>T. parva</i> TashAT cluster	135
Figure 3.12. Distribution of the G-box motif within intergenic regions of down-regulated genes	135

Figure 3.13. Cross species MEME analysis of the intergenic regions of <i>T. annulata</i> , <i>T. orientalis</i> and <i>T. parva</i> TA03110 orthologues	137
Figure 3.14. Cross species MEME analysis of the intergenic regions of <i>T. annulata</i> , <i>T. orientalis</i> and <i>T. parva</i> TA03110 orthologues	137
Figure 3.15. 5` intergenic sequence of TA03110	137
Figure 3.16. Domain alignment of all <i>T. annulata</i> ApiAP2 proteins using T-Coffee	145
Figure 3.17. Alignment of ApiAP2 domains of the four up-regulated <i>T. annulata</i> ApiAP2 genes (TA11145, TA13515, TA12015, TA16485) with their orthologues in <i>T. parva</i> and <i>T. orientalis</i>	147
Figure 3.18. ClustalW2 alignment of ApiAP2 domains from TA13515 orthologues in <i>T. parva</i> , <i>T. orientalis</i> and <i>P. falciparum</i>	148
Figure 3.19. ClustalW2 alignment of ApiAP2 domains from TA11145 orthologues in <i>T. parva</i> , <i>T. orientalis</i> and <i>P. falciparum</i>	148
Figure 3.20. ClustalW2 alignment of ApiAP2 domains from TA16485 orthologues in <i>T. parva</i> , <i>T. orientalis</i> and <i>P. falciparum</i>	148
Figure 3.21. ClustalW2 alignment of ApiAP2 domains from TA12015 orthologues in <i>T. parva</i> , <i>T. orientalis</i> and <i>C. parvum</i>	149
Figure 3.22. The C-box motif identified in the <i>T. annulata</i> Heat Shock gene family by MEME analysis	153
Figure 3.23. Negative correlation plots of ApiAP2 gene mRNA abundance and putative target genes possessing the motif bound by the AP2 domain	161
Figure 3.24. Positive correlation plots of ApiAP2 gene mRNA abundance and putative target genes possessing the motif bound by the AP2 domain	163
Figure 4.1. Alignment of the ApiAP2 domain from PF140633 (amino acids 63-123) to orthologues in five additional <i>Plasmodium spp.</i> and six apicomplexan species (<i>T. gondii</i> , <i>T. parva</i> , <i>T. annulata</i> , <i>B. bovis</i> , <i>Cryptosporidium hominis</i> , <i>C. parvum</i>)	175
Figure 4.2. <i>Plasmodium</i> ApiAP2 proteins binding the CACACA motif	179
Figure 4.3. Map of the GST- fusion vectors showing the reading frames and main features	183
Figure 4.4. Electrophoretic mobility shift assay (EMSA)	189

Figure 4.5. View of the top three hits of the TA13515 fold recognition results, including images of the protein models produced	198
Figure 4.6. View of the top three hits of the TA11145 fold recognition results, including images of the protein models produced	198
Figure 4.7. View of the top three hits of the TA16485 fold recognition results, including images of the protein models produced	199
Figure 4.8. View of the top three hits of the TA12015 fold recognition results, including images of the protein models produced	199
Figure 4.9. Secondary structure prediction of the TA13515 ApiAP2 domain	199
Figure 4.10. Secondary structure prediction of the TA16485 ApiAP2 domain	200
Figure 4.11. Secondary structure prediction of the TA11145 ApiAP2 domain	200
Figure 4.12. Secondary structure prediction of the TA12015 ApiAP2 domain	200
Figure 4.13. A 3-Dimensional structural prediction of four <i>T. annulata</i> ApiAP2 domains	201
Figure 4.14. GST-ApiAP2 domain fusion proteins generated in this study	202
Figure 4.15. EMSA assay performed with 0.7µg of purified GST-TA13515D and 20fmol of a 62bp 5`-biotin-labelled probe containing the GTGTACAC sequence motif	203
Figure 4.16. EMSA assay performed with 0.7µg of purified GST-TA13515D and 20fmol of biotin-labelled probe containing GTGTACAC motif; specific competitor and mutated probes	204
Figure 4.17. EMSA assay performed with 0.7µg of purified GST-TA13515D and 20fmol of biotin-labelled oligo probe containing the GTGTAC motif derived from the upstream region of the TA10735 gene and cold competitor probe	205
Figure 4.18. EMSA assay performed with 0.7µg of purified GST-AP2 domain fusion proteins and 20fmol of biotin-labelled oligo containing GTGTAC motif derived from the upstream region of the TA10735 gene	206
Figure 4.19. EMSA assay performed with 0.7µg of purified GST-TA13515D fusion protein and 20 fmol of biotin-labelled oligo containing the GTGTAC motif derived from the upstream region of TA10735 gene plus the addition of a PA sequence specific inhibitor (ISS-15)	207
Figure 4.20. EMSA assay performed with 0.7µg of purified GST-TA13515D fusion protein and 20 fmol of biotin-labelled oligo probe containing the	

	GTGTAC motif derived from upstream region of TA10735 gene with the addition of a control PA for non-specific binding (ISS-33)	208
Figure 4.21.	EMSA assay performed with 0.7µg of purified GST-ApiAP2 fusion proteins and 20 fmol of biotin-labelled oligo containing the CACAC motif derived from <i>Plasmodium</i> AP2 domain binding data	209
Figure 4.22.	EMSA assay performed with GST-TA11145D and labelled oligo probe containing 2x(A)CACAC(A) motifs derived from upstream region of the TA11145 ApiAP2 encoding gene	210
Figure 4.23.	GST-TA12015D expression in BL21 cells in TB medium	211
Figure 4.24.	Enhanced solubility of the GST-TA12015D in supernatant fraction by induction of protein expression in the presence of the different concentrations of Gly-Gly	211
Figure 4.25.	EMSA assay performed with labelled oligo probe containing C/G-box motif derived from the upstream region of the Tash1-like gene (TA03125) and increasing concentrations of GST-TA12015D	212
Figure 4.26.	EMSA assay performed with GST-TA16485D and biotinylated oligo probe containing the TCTATA motif	212
Figure 4.27.	EMSA experiment with PNE (D7, Day 0) and the labelled C-box oligo probe derived from the intergenic region of Tash1-like (TA03125) gene with and without cold probe competition	214
Figure 4.28.	EMSA experiment with PNE and the labelled C-box oligo probe derived from the intergenic region of Tash1-like (TA03125) gene	215
Figure 4.29.	Differential expression profiles of five ApiAP2 genes potentially binding CACACA motif in the D7 cell line	217
Figure 4.30.	<i>T. annulata</i> , <i>T. parva</i> and <i>T. orientalis</i> ApiAP2 alignment of domains that are likely to bind the CACACA motif	217
Figure 4.31.	Phylogenetic tree of the predicted <i>Theileria</i> and <i>Plasmodium</i> CACACA-binding ApiAP2 domains	218
Figure 4.32.	EMSA assay performed with D7 Parasite Enriched Nuclear Extract (PNE) from macroschizont (Day 0) and differentiating cultures (Day 7 and Day 9) in the presence of labelled oligo probe containing the ACACAC:CACACA tandem motifs derived from the upstream region of TA11145	219
Figure 4.33.	EMSA assay performed with D7 Parasite Nuclear Extract (PNE) from macroschizont (Day 0) in the presence of labelled oligo containing	

two (A)CACAC(A) motifs derived from the upstream region of TA11145 and specific and non-specific competitors	220
Figure 4.34. EMSA assay performed with D7 Parasite Nuclear Extract (PNE) and TA11145 upstream ACACA-CACACA labelled probe and mutated versions of this probe	222
Figure 4.35. Comparison of expression levels of TA11145 and TA07100 ((A)CACAC(A)-binding ApiAP2 genes) of particular time-points between D7 and D7B12 cell lines	225
Figure 4.36. EMSA assay performed with D7 and D7B12 Parasite Nuclear Extracts (PNEs) and the TA11145 upstream ACACAC-CACACA labelled probe	226
Figure 5.1. A regulatory network for the three macroschizont to merozoite up-regulated ApiAP2 genes	243

List of Tables

Table 1.1.	Characterisation of <i>Theileria annulata</i> , <i>T. parva</i> and <i>T. orientalis</i> genomes	31
Table 2.1.	List of hypothetical ApiAP2 proteins in <i>T. annulata</i>	55
Table 2.2.	Hypothetical transcription factors of <i>T. annulata</i> other than ApiAP2	56
Table 2.3.	Summary data on number of differentially gene expression obtained by pairwise comparison and RP analysis of microarray data sets of <i>T. annulata</i>	76
Table 2.4.	List of the thirty most up-regulated genes during differentiation from macroschizont to merozoite stage	78
Table 2.5.	List of the thirty most down-regulated genes during differentiation from macroschizont to merozoite stage	80
Table 2.6.	Semi-quantitative RT-PCR of up-regulated genes from macroschizont to merozoite stage	92
Table 2.7.	Semi-quantitative RT-PCR of down-regulated genes on tRNA from macroschizont to merozoite stage in vitro culture	93
Table 3.1.	Gene families enriched in 4C-box motif in their upstream regions	138
Table 3.2.	Statistical enrichment analysis of TCCCCAT motif within <i>T. annulata</i> intergenic regions	140
Table 3.3.	List of genes with 5C-box motif in their upstream region	142
Table 3.4.	Statistical enrichment analysis of TCCCCCAT motif within <i>T. annulata</i> intergenic regions	143
Table 3.5.	Orthologues of <i>T. annulata</i> up-regulated ApiAP2	146
Table 3.6.	<i>P. falciparum</i> ApiAP2 target gene motifs distribution in <i>T. annulata</i> upstream regions	151
Table 3.7.	Statistical enrichment analysis of (A/G)NGGGG(C/A) motif within <i>T. annulata</i> intergenic regions	154
Table 3.8.	Statistical enrichment analysis of GTGTAC motif within <i>T. annulata</i> intergenic regions	156
Table 3.9.	Statistical enrichment analysis of TCTA[C/T]A[A/G]A motif within <i>T. annulata</i> intergenic regions	157

Table 3.10.	Statistical enrichment analysis of [C/T][A/G]CACAC[C/T][A/T] motif within <i>T. annulata</i> intergenic regions	158
Table 3.11.	Number of DNA motifs (minimum of 4 core nucleotides) found in the upstream regions of the four macroschizont to merozoite up-regulated ApiAP2 genes	160
Table 4.1.	EMSA reaction components to study fusion AP2 protein-DNA interactions	190
Table 4.2.	EMSA reaction components to study parasite nuclear extract-DNA interactions	190
Table 4.3.	EMSA reaction components to study protein-DNA interactions in the presence of specific PA inhibitors	191
Table 4.4.	Characterisation of TaApiAP2 gene family	195
Table 4.5.	Properties of the recombinant fusion proteins produced in this study	202
Table 4.6.	Mutated biotinylated oligos used in EMSA experiment obtained from upstream region of TA11145 gene	221
Table 4.7.	Differential expression of TaApiAP2 genes between D7 and D7B12 cell lines	224

Acknowledgments

I would like to express my deepest gratitude to my main supervisor, Professor Brian Shiels. His guidance, encouragement and immense knowledge led me throughout this research degree. His frequent insights and patience with me were always appreciated. I would also like to thank my co-supervisors Dr William Weir for provided invaluable bioinformatic help and feedback on my work. And to Dr Jane Kinnaird, thank you for your experience in the laboratory techniques and making sure the research continued to progress towards a finished thesis. I have enjoyed the opportunity to watch and learn from her knowledge and experience. I appreciate my mentors' encouragement and willingness to provide consultation and guidance during the course of my study.

I would like to thank the funding source that made my PhD work possible. My project was part of Post-graduate training network for capacity building to control ticks and tick-borne diseases (POSTICK) and was funded by the European Commission Marie Curie fellowship. I would like to also thank University of Glasgow for accepting me to the PhD programme and providing laboratory to pursue the project.

I would like to thank the members of the *Theileria* group who have contributed immensely to my personal and professional time in Glasgow. I would like to thank especially Dr Laetitia Lempereur and Dr Zeeshan Durrani for their friendships as well as good advice and encouragement. I would like to also thank Sreerekha Sreedharan Pillai for her warm welcome and introduction to the laboratory when I joined the group.

A special thanks to my assessors: Professor Eileen Devaney and Professor Andy Waters for their valuable suggestions and encouragement during my studies. And Dr Abhinav Sinha for sharing his laboratory experience and providing contribution to my work from the *Plasmodium* point of view.

My time in Glasgow was enjoyable in large part also due to the many friends and groups that became a part of my life. I am grateful for time spent with flatmates

and friends who were also going through PhD experience and working long hours in the lab and at home. I am also thankful for the great group of PhD students from around the world who shared with me experience of studying ticks and tick-borne diseases within POSTICK project. Laetitia, without the experience and fun you provided during our work and travels it wouldn't be the same!

Lastly, I would like to thank my family for all their love and encouragement. For my parents who supported me in all my pursuits. And most of all for my loving, encouraging and patient boyfriend Andy whose support during the final stages of this PhD was very appreciated. Thank you.

Marta Pieszko

List of Abbreviations and Symbols

α	alfa
ACDC	AP2-coincident C-terminal domain
AP2	Apetala2, family of transcription factors
AP2-G	Apetala2 factor triggering gametogenesis in <i>Plasmodium</i>
AP2-Sp	Apetala2 factor in <i>Plasmodium</i> sporozoites
AP2-O	Apetala 2 factor in <i>Plasmodium</i> ookinetes
AP2IX-9	<i>Toxoplasma</i> Apetala2, repressor of bradyzoite development
ApiAP2	Apicomplexan Apetala 2 family of transcription factors
AT	adenine thymine
ATG	translation start codon; Adenine, Thymine, Guanine
β	beta
BL20	uninfected bovine leucocytes cell line
bp	base pair
BSA	bovine serum albumin
BW720c	Buparvaquone
cDNA	complementary deoxyribonucleic acid
CO ₂	carbon dioxide
°C	Degrees celsius
Ct	cycle threshold
CHD1	chromodomain helicase DNA binding protein 1
$\Delta\Delta$ Ct	delta delta, cycle threshold
D7	<i>T. annulata</i> infected differentiating clonal cell line
D7B12	<i>T. annulata</i> infected non-differentiating clonal cell line
DAPI	4',6-diamidino-2-phenylindole, dihydrochloride
DABCO	1,4-diazabicyclo[2.2.2]octane
DB	data base
DBD	DNA binding domain
DEPC	Diethylpyrocarbonate
dH ₂ O	distilled water
DMSO	dimethyl sulfoxide
DNase	Deoxyribonuclease
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotidetriphosphate

3D	three dimensional
EBNA	Epstein-Barr Nuclear Antigen
ECF	East Coast Fever
EDTA	Ethylenediaminetetracetic acid
EMSA	Electrophoretic Mobility Shift Assay
EST	expressed sequence tag
EtBr	Ethidium bromide
FBS	foetal bovine serum
FDR	False Discovery Rate
g	gram
GC	guanine cytosine
GCN5	member of Histone acetyltransferase family
Gly-Gly	Glycine-Glycyl
GST	glutathione S-transferase
H	hour
HAT	histone acetyltransferase
HCl	Hydrogen Chloride/ Hydrochloric acid
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HMGB	High-mobility group box proteins
HSP	Heat Shock Protein
Hsp70	Heat Shock Protein, 70 kDa
Hsp90	Heat Shock Protein, 90 kDa
IDC	intraerythrocytic developmental cycle
IFAT	Immunofluorescent antibody test
Ig	Immunoglobulin
IGR	intergenic region
IPTG	isopropyl B-D thiogalactoside
kb	Kilobase or 1000bp
KCl	Potassium chloride
kDa	Kilodalton
KH ₂ PO ₄	Potassium dihydrogen orthophosphate
l	litre
LB	Lysogeny Broth
LT	labeled target
M	molar

MEME	Multiple Expectation Maximization for Elicitation of Motifs
MeR	regulator of merozoite gene expression
mg	milligram
MgCl ₂	Magnesium Chloride
min	minute
ml	millilitre (10 ⁻³)
mM	millimolar
MOPS	4-Morpholinepropanesulfonic acid
mRNA	messenger ribonucleic acid
Myb	DNA binding domain of Myb transcription factor family
μ	micro (10 ⁻⁶)
μg	microgram
μl	microlitre
μM	micromolar
NaCl	sodium chloride
Na ₂ HPO ₄	disodium hydrogen phosphate anhydrous
N	number of samples
NaOH	Sodium hydroxide
NF-κB	nuclear factor-kappa B
ng	nanogram (10 ⁻⁹)
nm	nanometer
NP-40	nonyl phenoxy polyethoxy ethanol
OD ₆₀₀	optical density measured at wavelength of 600nm
oligo	oligodeoxyribonucleotide
O/N	over-night
ORF	Open Reading Frame
p	p value; statistical significance
PAGE	Polyacrylamide gel electrophoresis
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PFA	Paraformaldehyde
Pfu	proof-reading polymerase
pGex	bacterial vector for expressing GST-fusion proteins
PMSF	phenylmethanesulfonyl fluoride
PNE	parasite-derived nuclear extract

Poly (dIdC)	Poly(deoxyinosinic-deoxycytidylic) acid
RP	Rank Product
qRT-PCR	Quantitative reverse transcription PCR
RNA	Ribonucleic acid
RNase	Ribonuclease
RPMI	Roswell Park Memorial Institute basal culture medium
rRNA	ribosomal ribonucleic acid
RT	Reverse transcriptase
RT-PCR	Reverse transcription PCR
rpm	revolutions per minute
SDS	Sodium dodecyl sulphate
SDS-PAGE	sodium dodecyl sulphate polyacrylamide gel electrophoresis
sec	second
SEM	Standard error of the mean
SOC	Super Optimal broth with Catabolic repressor
SPAG-1	<i>T. annulata</i> sporozoite surface antigen
SuAT1	<i>T. annulata</i> encoded host-nuclear protein (TA03135)
SVSP	sub-telomeric variable secreted protein family
Ta9	<i>T. annulata</i> hypothetical protein (TA15705)
Ta	Annealing temperature
Tams1	<i>T. annulata</i> merozoite surface antigen (TA17050)
Tash-a	<i>T. annulata</i> schizont host-nuclear protein (TA03110)
TashAT3	<i>T. annulata</i> schizont host-nuclear protein (TA20082)
TashAT1	<i>T. annulata</i> schizont host-nuclear protein (TA20085)
TashHN	<i>T. annulata</i> schizont host-nuclear protein (TA20090)
TashAT2	<i>T. annulata</i> schizont host-nuclear protein (TA20095)
TaSP	<i>T. annulata</i> surface protein (TA17315)
Taq	<i>Thermus aquaticus</i> polymerase
TAE	Tris-Acetate EDTA
TB	Terrific Broth
TBE	Tris-Borate EDTA
TBD	Tick-borne disease
TBS	Tris-Buffered Saline
TF	transcription factor
Tm	melting temperature

Tris	Tris(hydroxymethyl) amino methane
tRNA	transfer RNA
TSS	translational start site
UV	Ultraviolet
UPGMA	Unweighted Pair Group Method with Arithmetic Mean
UTR	Untranslated region
V	Volt
v	Volume

Chapter One

Introduction

1 Introduction

1.1 Introduction

Apicomplexan parasites are protozoan pathogens of humans and domestic and wild animals and represent an on-going threat to livestock productivity across much of the world. *Theileria annulata* is an intra-cellular protozoan parasite that causes tropical theileriosis, an economically important tick-borne disease (TBD) of cattle. It is found in Southern Europe, largely around the Mediterranean coast, the Middle East, North Africa and much of central and East Asia including India (Purnell et al., 1978, Lu and Yin, 1994) (Figure 1.1). The geographical distribution of *T. annulata*, and hence tropical theileriosis, is mainly determined by the location and biology of its vector, ticks of the *Hyalomma* genus (Robinson, 1982). A suitable climate (i.e. warm temperatures and high humidity) is required for tick survival, questing and trans-stadial transmission of the parasite. *Theileria annulata* has been reported as transmitted by *H. detritum* and *H. truncatum* in Africa, *H. dromedarii* in central Asia, *H. asiaticum* in China and *H. anatolicum* and *H. marginatum* in India (Flach and Ouhelli, 1992; Bouattour et al., 1996; Islam et al., 2006; Taylor et al., 2007; Meng et al., 2014).

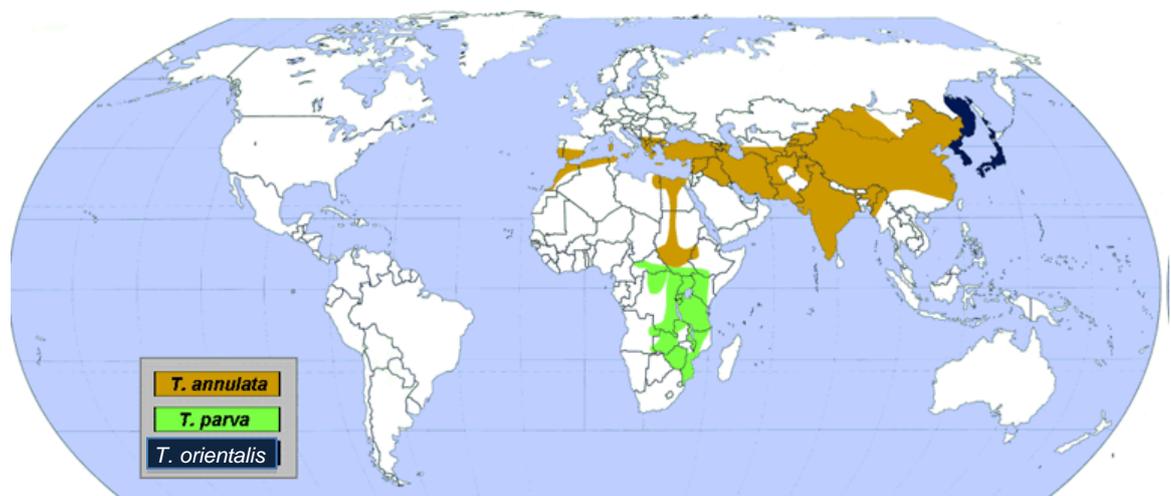


Figure 1.1 Distribution of major *Theileria* species infecting cattle

Theileria annulata is found in Southern Europe, North Africa and sub-Saharan regions of Sudan and Eritrea, the Middle East, India and Southern China (highlighted in brown). *Theileria parva* is found in south-east Africa (highlighted in green). *Theileria orientalis* is found in far-eastern Asia including Japan and Korea (highlighted in blue).

(www.theileria.org/ahdw/pictures/largemap.gif)

Domestic cattle (*Bos taurus* and *Bos indicus*) are susceptible to *T. annulata* infection as are yaks (*Bos grunniens*), water buffalo (*Bubalus bubalis*) and camels (*Camelus dromedarius*) (Robinson, 1982; Nassar, 1992; Dumanli et al., 2005). *Theileria annulata* sporozoites can be transmitted to goat and sheep and cause mild febrile response, however limited experimental studies indicate that schizonts and piroplasms are not produced in these host species (Leemans et al., 1998; Yin et al., 2003) and therefore the parasite cannot complete its life-cycle. Infection by *T. annulata* parasites results in enormous losses in production (beef and milk) (Gharbi et al., 2006), with high mortality in susceptible animals (Branco et al., 2010). Other major *Theileria* species of domestic cattle include *T. parva* in sub-saharan Africa, and highly-related phylogenetically *T. buffeli*/*T. orientalis* group (Sugimoto et al., 1991) with *T. orientalis* in Japan and Korea and *T. buffeli* over much of the world and considered a recently emerging pathogen in Australia (Islam et al., 2011). In sheep and goats, *Theileria lestoquardi* causes severe disease and is associated with high morbidity and mortality (Brown, 1990). Many other species of *Theileria* exist, particularly in sub-saharan Africa, and are carried by domesticated and wild ruminants with, in general, little or no clinical signs.

1.2 *Theileria* and closely-related Apicomplexa species

The phylum Apicomplexa comprises more than 5,000 species, which display a variety of life-cycle adaptations. However, these protozoa share fundamental aspects of their biology, i.e. differentiation through sequential life-cycle stages, amplification within the host cell and onward transmission to new hosts. At the present time little is known about the molecular mechanisms that directly regulate these transition events. Apicomplexan species are responsible for many important human and animal diseases such as mosquito-borne malaria (caused by *Plasmodium* species) and tick borne disease (TBD) such as babesiosis (caused by *Babesia*), East Coast fever (caused by *Theileria parva*) and tropical theileriosis (caused by *T. annulata*). *Toxoplasma gondii* and *Cryptosporidium parvum* parasites cause toxoplasmosis and cryptosporidiosis respectively, however they do not require vector organisms as they are transmitted directly from one

mammalian host via resistant stages in the environment and, in the case of *T. gondii*, via an intermediate host infected with a tissue cyst.

Theileria parasites show similarity at the cellular and molecular level to other Apicomplexan genera, in particular *Babesia* and *Plasmodium* (Figure 1.2). Every member of this phylum shares the same infectious stage, the sporozoite, which possesses a unique plastid-type organelle involved in energy metabolism (Lim and McFadden, 2009), the apicoplast; and the apical complex structure involved in host cell penetration. *Theileria annulata* and *T. parva* are phylogenetically very closely related species, which share the ability to transform and immortalise their host cells and induce un-controlled proliferation (Hulliger, 1965; Irvin and Morrison, 1987; Baylis et al., 1995; Adamson et al., 2000). Despite their close relationship, they display different host cell tropism and are transmitted by different tick species. *Theileria parva* has a more limited geographical range, as it is transmitted by the *Rhipicephalus appendiculatus* tick and is found only in Eastern and South Africa, south of the Sahara desert. It shows higher virulence, in terms of disease severity, than *T. annulata* (Norval et al., 1992). *Theileria lestoquardi* is a pathogenic species causing malignant theileriosis in sheep and goats (Uilenberg, 1981). Phylogenetically, it is highly similar to *T. annulata* and is also capable of reversible host cell transformation (Dobbelaere and Heussler, 1999; Pain et al., 2005). Non-transforming species of the genus, which tend to be less pathogenic, include parasites within the *T. buffeli*/*T. orientalis* complex (Onuma et al., 1997).

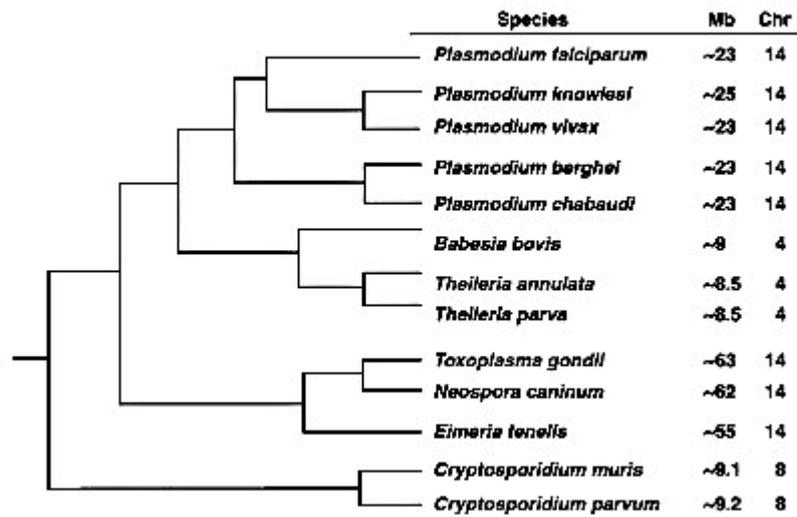


Figure 1.2 Relationship between Apicomplexa species and their genome sizes (Kissinger and DeBarry, 2011). Phylogenetic tree of Apicomplexa parasites. Approximate genome sizes are shown in Mb. Genomes of Apicomplexa are organised in 4 to 14 linear chromosomes.

The complete genome of *T. annulata* has been sequenced and is relatively small (~8.5 Mb) in comparison to that of other Apicomplexa, e.g. *T. gondii* (~63 Mb), *P. falciparum* (~23 Mb) and *C. parvum* (~9.2 Mb) (Figure 1.2 and Table 1.1). *Theileria annulata* has four chromosomes that range in size from 1.9 to 2.6 Mb (Pain et al., 2005). In contrast, the core *Plasmodium sp.* Genome spans 14 chromosomes (Lau, 2009).

Nuclear genome feature	<i>T. orientalis</i>	<i>T. annulata</i>	<i>T. parva</i>
Size (Mbp)	9.0	8.4	8.3
No. of chromosomes	4	4	4
Total G+C content (%)	41.6	32.5	34.1
No. of protein-coding genes	3,995	3,792	4,035
% of genes with introns	78.3	70.6	73.6
Mean gene length (bp)	1,861	1,606	1,407
% Coding	68.6	72.8	68.4
Mean intergenic length (bp)	390	396	402
% G+C composition of exons	44.5	37.6	35.9
% G+C composition of intergenic regions	35.2	22.5	24.9
% G+C composition of introns	38.1	22.2	23.6
No. of tRNA genes	47	47	47
No. of 5S rRNA genes	3	3	3
No. of 5.8S, 18S, and 28S rRNA units	2	2	2
Mitochondrial genome size (kb)	2.5	6	6
Apicoplast genome size (kb)	26.5	NA	39.5
Gene density ^a	2,249	2,202	2,059

Table 1.1 Characterisation of *Theileria annulata*, *T. parva* and *T. orientalis* genomes (Hayashida et al., 2012).

The *T. annulata* genome shows similarity to *T. parva* and *T. orientalis* both in terms of gene content and synteny and contains approximately 4,000 protein-coding genes, 47 tRNA genes and 5 ribosomal RNA genes (Pain et al., 2005; Hayashida et al., 2012). Some apicomplexan genomes are more AT-rich than others, with *P. falciparum* having an 80 % AT content (Gardner et al., 2002) compared to 67.5 % for *T. annulata* (Hayashida et al., 2012) and 50% for *Toxoplasma* (Gissot et al., 2008). The small size of the *Theileria* genome in comparison to most other eukaryotes has been attributed, in part, to the absence of genes involved in a range of metabolic processes. This includes genes involved in the synthesis of purines, polyamines, fatty acids and porphyrin. However, the presence of a number of key genes indicates that *Theileria* are capable of glycolysis and probably have a tricarboxylic acid cycle (Gardner et al., 2005).

1.3 Economic importance and pathogenesis of tropical theileriosis

The first description of tropical theileriosis, then called tropical piroplasmiasis, in cattle was in Transcaucasia, a region comprising today's Armenia, Azerbaijan and Georgia and the borders of Turkey and Iran (Dschunkowsky and Luhs, 1904). The parasite, of a similar shape to *T. parva* erythrocytic forms, was initially named *Piroplasma annulata*, but included in the genus *Theileria* after discovery of the multinucleate intracellular macroschizont stage (Bewttencourt et al., 1907). *Theileria annulata* and *T. parva* are closely related parasites and give rise to similar clinical syndromes (Irvin and Morrison, 1987; Allsopp and Allsopp, 2006), however *T. annulata* preferentially infects B cells (lymphocytes) and cells of myeloid lineage, while *T. parva* prefers T cells (lymphocytes) (Glass et al., 1989; Spooner et al., 1989).

The incubation period for tropical theileriosis lasts between approximately one and three weeks (Neitz et al., 1957). The severity of disease is influenced by the breed and immune status of the host species and is associated with the number of sporozoites inoculated by the tick (Samantaray et al., 1980; Preston et al., 1992). Both the macroschizont infected leukocyte, and the piroplasm infected erythrocyte, contribute to the pathogenesis of tropical theileriosis (Irvin and

Morrison, 1987). The most prominent clinical signs of the acute form of disease are pyrexia (41°C), lymphadenopathy, anaemia, jaundice, tachycardia and heart failure (Irvin et al., 1981; Preston et al., 1992). During the advanced third phase, severe anaemia and leukopenia occurs and this is due to widespread lymphocytolysis (Preston et al., 1992; Forsyth et al., 1999; Singh et al., 2001). Macroschizont-infected cells can be detected in biopsy smears of lymph nodes of live cattle approximately 7 to 28 days post-infection (Young et al., 1977). Infiltration of the lungs by infected cells can cause severe pulmonary oedema and death may ensue within 3 to 4 weeks of infection (Irvin and Morrison, 1987).

Theileria-infected cells show many characteristics of neoplastic cells (that undergo abnormal growth, not coordinated with other normal cells), including migration and proliferation in lymphoid and non-lymphoid tissues and the ability to form lesions similar to lymphosarcoma (Dobbelaere and Heussler, 1999). Cattle that have recovered from clinical tropical theileriosis, or were sub-clinically infected, usually remain parasite positive and carry *T. annulata* for an extended period of time. In addition in sub-acutely infected carriers, cows' fertility level may be reduced and milk production can become severely impaired (Dobbelaere and Heussler, 1999). In countries where tropical theileriosis is endemic, exotic *Bos taurus* cattle, and cross-breed Holstein-Friesians, are more susceptible to *T. annulata* infection with a mortality rate around 40-60 %, in contrast to indigenous *Bos indicus* cattle with mortality of only 5 % (Dyer and Tait, 1987; Brown, 1990; Bakheit and Latif, 2002; Glass et al., 2005). The natural resistance of indigenous cattle to theileriosis may be due to adaptive immunity and host/parasite co-evolution. Through immune recognition by cattle and the establishment of a carrier state that avoids clearance of the parasite a form of endemic stability is created (Antia et al., 1996), characterised by high immunity in adult cattle and absence of clinical disease in calves despite a high level of infection pressure. In endemic areas ticks usually take in a minimal dose of *T. annulata* from carrier cattle, that maintains infection, and may help in establishing endemic stability of theileriosis (epidemiological state in which despite high level of infection, the clinical disease is sporadic) by limiting the infective dose for naive susceptible animals (Haque et al., 2010).

Approximately 300 million cattle in endemic regions are at risk from *T. annulata* infection (Norval et al., 1992; Gharbi et al., 2006). In India alone, tropical theileriosis is a major health problem for up to ten million cattle and has been

associated with estimated economic losses up to 800 US million dollars per year (Griffiths and McCosker, 1990; Brown, 1997). These losses include death of animals, decreased productivity and the cost of disease control measures. Many of the non-transforming *Theileria* species (e.g. *T. mutans*) are carried without obvious clinical signs in domesticated and wild ruminants, although some can cause fever, anaemia or anorexia and increase the severity of East Coast fever or tropical theileriosis.

1.4 *Theileria annulata* control strategies

Several methods for the treatment and control of *T. annulata* infection have been used in endemic regions around the world. Tropical theileriosis can be controlled by immunisation with a live attenuated (of reduced virulence) vaccine (Pipano, 1981), obtained by prolonged cultivation of macroschizont-infected cells (Hall et al., 1999). Additionally application of acaricides or chemotherapy is advised (Dolan, 1989; Boulter and Hall, 1999). Vaccination against tropical theileriosis has been a focus of research for more than half a century and vaccination programmes have been deployed, with varying degrees of success, in a number of countries. Studies on immune responses to *T. annulata* suggested that recombinant vaccines should at least include antigens against sporozoites and macroschizonts (Preston et al., 1999). Immunisation with recombinant sporozoite surface antigen: SPAG-1 (*T. annulata*) and p67 (*T. parva*) induced a degree of protection against the respective *Theileria* species. Cross-species protection was also observed (Hall et al., 2000). Unfortunately, field trials against ECF did not support commercialisation and deployment of a recombinant vaccine based on this antigen alone (Morzaria et al., 2000). Vaccination with attenuated, cultured macroschizonts did not result in transmission to *Hyalomma* ticks, however carrier status developed in vaccinated calves (Gubbles et al., 2000). Stronger immunity was obtained from blood-delivered macroschizonts compared to culture-derived schizonts (Pipano and Shkap, 2000). Duration of protection does not appear to be lifelong, however safe and effective in large endemic areas. Attempts at developing a subunit vaccine based on the merozoite surface antigen (Tams1) were also undertaken, with good protection obtained against a blood challenge but only partial protection against a sporozoite

challenge (d'Oliveira et al., 1997; Gubbels et al., 2000b). A cocktail of SPAG-1 and Tams1 was recently shown to generate a synergistic response against sporozoite challenge (Gharbi et al., 2011) and similar results have been obtained by combining SPAG-1 with a live attenuated vaccine in Tunisia (Darghouth et al., 2006). Recent attempts have been made to test recombinant vaccine comprised of antigens of the tick, *H. a. anatolicum*, or parasite *T. annulata* (Jeyebal et al., 2012). Evidence of protection was obtained with the highest immune response generated against the tick-derived rHaa86 antigen, which effectively protected calves against tick infestation. Upon challenge with *T. annulata* infected ticks, parasite derived SPAG-1+rTaSP antigen immunized calves died, while 50% of calves immunised with rHaa85 survived. It was concluded that rHaa85 should be considered as a candidate for a vaccine against both *H. a. anatolicum* and *T. annulata*.

Transformation induced by *Theileria* is a reversible process and the addition of theilericidal drugs has been shown to stop proliferation of cells. Several chemotherapeutics have been used to treat theileriosis, such as oxytetracyclines (Bansal and Sharma, 1986), parvaquone (McHardy and Morgan, 1985) or halofuginone lactate (Mehlhorn and Raether, 1988). However, the most effective drug in use worldwide is buparvaquone (BW720C), a second generation hydroxynaphthaquinone which has the ability to cure cattle infected with *T. annulata* and *T. parva* (Dolan et al., 1984; McHardy et al., 1985; Hawa et al., 1988; Ngumi et al., 1992; Singh et al., 1993). Unfortunately, therapeutics are most efficacious if deployed in the early stages of disease and their high cost precludes widespread use in developing countries where the disease is endemic (Morrison and McKeever, 2006; Osman and Al-Gaabary, 2007). In addition, emerging drug resistance and delivery against only the parasite life stages in the bovine host limits the effectiveness of the agents (Mallick et al., 1978; Schein and Voight, 1979; Singh et al., 1993; Mhadhbi et al., 2010). A combined strategy of vector and parasite control is therefore required to combat tropical theileriosis. To improve current measures of control, one potential approach is to develop a system that targets parasite stage differentiation, thus impairing the development of life-cycle stages responsible for the pathogenesis in cattle or transmission to the vector. This would require greater understanding of the mechanism that controls these events. It is likely that existing drugs operate to

modulate stage differentiation of *Theileria* and other apicomplexans (Shiels et al., 1998).

1.5 Life-cycle of *T. annulata*

Apicomplexan species demonstrate a variety of complex life-cycles most if not all involving a step that allows transition from asexual to sexual stages. *Theileria* parasites undergo a number of differentiation stages that are typical of the Apicomplexa, including gametocytogenesis, sporogony and merogony (Figure 1.3). The life-cycle can be divided into two main phases: stages in the tick (vector) and stages within the mammalian host. To be transmitted between mammalian hosts or between host and vector, apicomplexan parasites must undergo differentiation into morphologically distinct forms adapted to survive in different environments. For *Plasmodium* this involves transition from the intraerythrocytic developmental cycle (IDC) stages to the gametocytes and sexual determination, while for *T. annulata* differentiation from the intracellular macroschizont to extracellular merozoites is followed by invasion of erythrocytes and formation of the piroplasm stage that allows transmission to the tick. An array of host-parasite interactions underpin the ability of macroschizont-infected cells to survive, proliferate and disseminate. However, if the life cycle is to progress and new host to be infected stage differentiation must occur. Differentiation is also critical to establish infection, and disease, following inoculation of the sporozoite. Thus, stage differentiation allows regeneration and expansion of parasite populations and thus can be considered a target for controlling the transmission of parasites to their vector and preventing infection from reaching clinical outcome. Blocking differentiation events in the tick also has the potential to moderate and eliminate transmission of the pathogen to susceptible livestock.

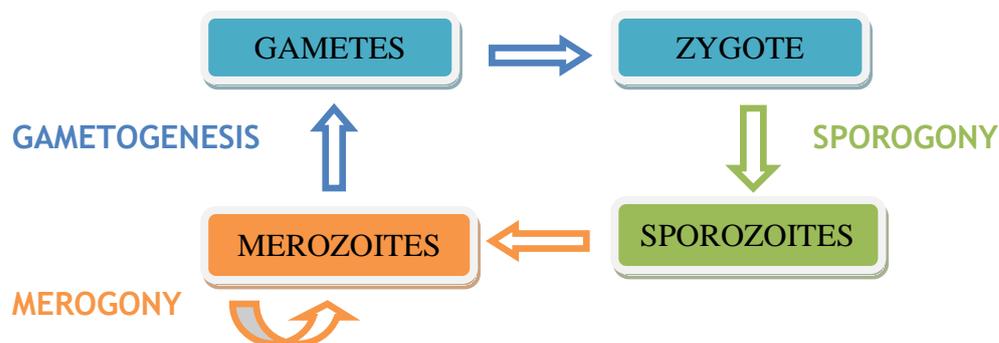


Figure 1.3 Generalised apicomplexan life-cycle

1.5.1 Bovine stages of the *T. annulata* life-cycle

There are a number of distinct stages within the bovine phase of the life-cycle, each of which are asexual. The bovine phase of the *T. annulata* life-cycle (Figure 1.4) begins with inoculation of parasite sporozoites with the saliva of the feeding tick. Sporozoites are oval shaped bodies of approx. 1µm in length that invade leukocytes (B lymphocytes or monocytes) by discharging their rhoptries and micronemes, specialised secretory organelles involved in facilitating host cell invasion, to lyse the host cell plasma membrane (Fawcett et al., 1982, Shaw et al., 1991) that initially surrounds the invading sporozoite. This results in a free parasite (trophozoite) surrounded by host cell microtubules located in the perinuclear region of the host cell. After 2-3 days in the cytoplasm, the trophozoite differentiates into the multinuclear macroschizont (Fawcett et al., 1982, Shaw et al., 1991). It is the macroschizont stage that possesses the ability to immortalise the host cell (Baumgartner et al., 2000). This is achieved by activation of host cell transcription factors and modulation of host cell gene expression (reviewed by Shiels et al., 2005; Durrani et al., 2012; Kinnaird et al., 2013). The parasite associates closely with the microtubules at the host cell spindle in order to ensure distribution of the parasite to each daughter leukocyte during cytokinesis (Hullinger et al., 1964; Mehlhorn and Schein, 1984). Proliferating macroschizont-infected leukocytes give rise to a large number of infected cells, initially in the lymph nodes draining the site of tick attachment. They then disseminate to the rest of the lymphoid system and can be found in the bloodstream and a variety of non-lymphoid tissues and organs (Forsyth et al., 1999; Glass, 2001). Dissemination of infected cells and the host reaction to them are responsible for the much of the pathogenesis and clinical signs of the

disease, which has been described as being similar to a neoplastic condition (Swan et al., 2003; Lizundia et al., 2006; Chaussepied et al., 2010; Cock-Rada et al., 2012). The ability to induce the host cell to divide is unique to *Theileria* and is almost certainly linked to the mechanism of proliferation and its ability to co-opt the host cell mitotic spindle (von Schubert et al., 2010). *Plasmodium* and *Toxoplasma*, for example can both significantly modulate nucleated host cells, but reside within a parasitophorous vacuole and divide within this vacuole independently of the host cell (Shaw, 1997).

Macroschizonts differentiate to become uni-nucleated extracellular invasive forms called merozoites through a process known as merogony (Jarret et al., 1969; Shaw et al., 1992). This is a major point of the differentiation process in the bovine phase of the parasite life-cycle as it involves the switch in production of macroschizont to merozoite proteins, formation and assembly of merozoite nuclei, microneme and rhoptry organelles and is marked by major changes to parasite cellular morphology and antigenic profile (Mehlhorn and Shein, 1984; Shiels et al., 1992). During the initial phase of merogony proliferation of the infected host cell slows down (Shiels et al., 1992), the parasite syncytium enlarges to generate host cells containing a large number of mature merozoites (Shaw et al., 1992). In contrast to *T. annulata* and *T. parva*, merozoite production of *T. orientalis* occurs to a significant degree in erythrocyte cells (Conrad et al., 1985). *Theileria orientalis* also undergoes merogony in host leukocytes, but is incapable of causing host-cell transformation to induce proliferation (Nene et al., 2000). Thus immortalised *T. orientalis*-infected cell lines cannot be established *in vitro*, and *in vivo* the host cell increases substantially in size with macroschizonts undergoing continuous enlargement from 4 to 8 days before multiple merozoites are generated and released (Hayashida et al., 2012). *Theileria* parasites, similar to *Babesia* parasites, are thought to undergo sexual differentiation with formation of a zygote in the tick vector (Mehlhorn and Schein, 1984). In *Plasmodium* species, the sexual erythrocytic phase starts with invasion of committed merozoites derived from the previous round of asexual schizogony. The resulting ring-form parasites differentiate to trophozoites (Garcia et al., 2008) that undergo the process of gametocytogenesis, rather than undergo another IDC. In a similar manner the *Cryptosporidium parvum* life-cycle involves both asexual and sexual stages with infection initiated by the ingestion of oocysts. Oocysts release sporozoites,

which in turn invade intestinal epithelial cells and undergo asexual amplification via schizogony and merozoite production. After several rounds of schizogony, the merozoites become committed to give rise to gametocyte production (Chen et al., 2002). Thus differentiation of the parasite to the merozoite stage and subsequently to the sexual stages is common across distantly related Apicomplexa genera and it is reasonable to speculate that related underlying molecular mechanisms control these life cycle events.

Following destruction of the leukocyte membrane and budding from the syncytial macroschizont, uni-nucleated merozoites are released into the host bloodstream. Merozoites proceed to invade the erythrocytes and mature into piroplasms within 8-10 days of sporozoite inoculation. Piroplasms are rod or oval-shaped bodies surrounded by a single-layered plasma membrane, which lie freely in the cytoplasm of the erythrocyte (Mehlhorn and Schein, 1984). They have been considered by some authors to be the equivalent of gametocytes (Mehlhorn and Schein, 1984). The life-cycle in the bovine host is completed with the ingestion of piroplasm infected erythrocytes by a feeding tick (Norval et al., 1992).

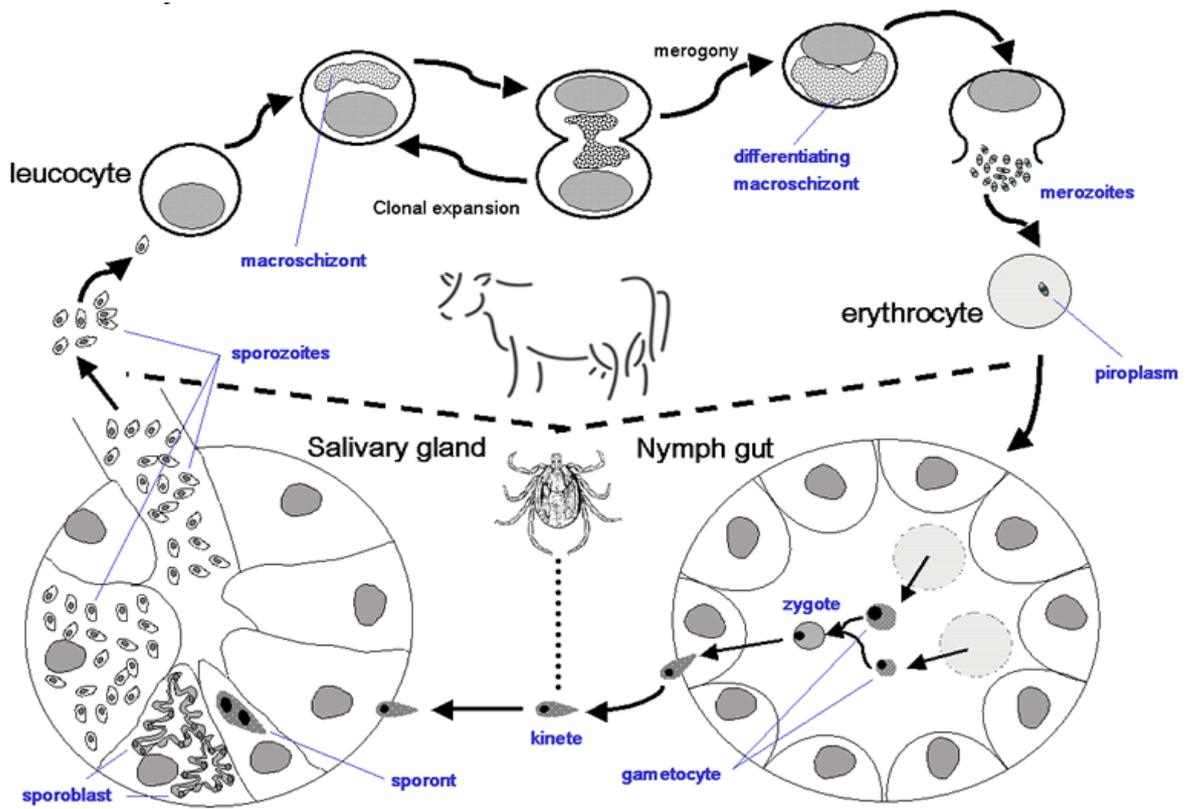


Figure 1.4 The life-cycle of *T. annulata*
 (www.theileria.org)

1.5.2 Tick stages of the *T. annulata* life-cycle

The tick is infected by feeding on a bovine infected with *Theileria* piroplasms in its red blood cells. These enter the tick gut lumen where gametogenesis and fertilisation to produce a zygote occur (Gauer et al., 1995). Piroplasms develop into spindle-shaped bodies, micro- or macrogamonts, which then transform into uni-nucleated micro- and macrogametes (Schein et al., 1975). Fusion of micro- and macrogametes takes place and immotile zygotes are formed (Schein et al., 1975; Levine, 1985).

After 12-15 days, the *T. annulata* zygote differentiates into a motile elongated body called a kinete (Mehlhorn and Schein, 1984), which is a polyploid life-cycle stage. In *T. parva*, it has been proposed that the tetraploid zygote undergoes a two-step meiosis in the tick gut epithelium and the DNA complement is reduced to a haploid level (Gauer et al., 1995). Following the tick moult, the kinete invades the salivary glands, undergoes multiple nuclear divisions and forms a complex syncytium called the sporoblast (Schein and Friedhoff, 1978; Mehlhorn and Schein, 1984). Following 3 or 4 days of feeding on a new host, a large number of sporozoites are produced in the salivary glands of the tick by the process of asexual nuclear division and budding, known as sporogony (Singh et al., 1979; Reid and Bell, 1984). Approximately 40,000 sporozoites are generated from each infected gland acinus (Young et al., 1992). Sporozoites are released into the tick saliva and delivered into the new bovine host during feeding.

1.6 A stochastic model of stage differentiation in *T. annulata*

Although the steps involved in differentiation vary among apicomplexan parasites, some common characteristics of these processes have been observed. These include alteration of environmental temperature during the life-cycle as they move from one vector/host to another (Hullinger et al., 1966; Soete et al., 1994), and a reduction of parasite proliferation status (Bruce et al., 1990; Bohne et al., 1994; Reuner et al., 1997; Shiels et al., 1992). Changes to heat shock gene expression levels were also proposed to be involved in the regulation of

differentiation processes in other protozoa: e.g. *Leishmania* (Van Der Ploeg et al., 1985). Signalling mechanisms have been frequently proposed to operate in stage differentiation of protozoa. Signalling events controlling gametocytogenesis has been poorly characterised in *P. falciparum*, however recent discovery that *P. falciparum* infected red blood cell-cell communication through exosome-like vesicles act as a messenger for the induction of differentiation to the sexual forms suggests that signalling might be involved in this process (Mantel et al., 2013; Regev-Rudzki et al., 2013). Moreover, the cGMP-dependent protein kinase was recently identified as a mediator of this process (McRobert et al., 2008). However, contradictory to a signalling event, evidence has been generated which indicates that differentiation in protozoa is an asynchronous/stochastic process and is associated with stoichiometric changes in the level of molecular factors regulating gene expression (Shiels et al., 1994; Soete et al., 1994). A stochastic (with random probability) differentiation process has previously been described for *Trypanosoma brucei* (Reuner et al., 1997), *T. gondii* (Soete et al., 1994), *T. annulata* (Shiels et al., 1992), *Plasmodium* (Carter and Miller, 1979; Bruce et al. 1990), *Eimeria* (McDonald and Rose, 1987) and *Leishmania* parasites (Sacks and Perkins, 1894). It has been also shown in *T. annulata*, *P. falciparum* and *T. gondii* that certain drugs which inhibit DNA synthesis can increase differentiation potential of these parasites (Shiels et al., 1997; Bohne et al., 1994; Ono et al., 1993). Other studies observed that a number of protozoa show evidence of early expression of stage-specific markers in the preceding stage, before the differentiation event occurs (Contreras et al., 1985; Soete et al., 1994; Shiels et al., 1994; Abrahamson et al., 1995; Reuner et al., 1997), implying to a degree that the event is predetermined. For the *Theileria* model an initial low-level increase in expression of merozoite genes could be reversed up to a certain point but then became irreversible (Shiels et al., 1994). These results indicated a point of commitment determined when the differentiation event becomes irreversible. Thus, Shiels et al. (1994) proposed that differentiation to the merozoite in *T. annulata* operates on the basis of factors that regulate gene expression reaching a commitment threshold, which, once reached, results in an irreversible switch over from macroschizont to merozoite gene expression (Figure 1.5). This hypothesis was supported by the switch to irreversible high-level Tams1 gene expression, a major antigenic component of the merozoite surface, together with loss of expression of

macroschizont specific antigens (Shiels et al., 1994). Importantly macroschizonts express the Tams1 gene at a detectable level and a small reversible increase in expression was described during 2-4 days after the cell line being put at 41 °C (Shiels et al., 1999). In addition, cells with an attenuated differentiation phenotype were unable to significantly increase Tams1 gene expression (Shiels et al., 1999), implying that the putative regulator of merozoite gene expression was unable to reach the threshold required for the switch to merogony.

To explain low level expression of the Tams1 gene in the preceding stage it was proposed that either merozoite factors function at a low level in the macroschizont or there is cross-recognition of merozoite promoters by factors that control macroschizont gene expression. Then from experiments that altered the level of DNA synthesis relative to protein synthesis (Shiels et al., 1997) it was postulated that following placement at 41 °C (temperature of fever) a general increase in factor production over DNA template leads to a preferential increase in regulators of merozoite gene expression via a positive feedback loop until full auto-regulation and the commitment point of differentiation is achieved (Shiels et al., 1999).

Further investigation of Tams1 gene expression demonstrated regulation at the transcriptional level and provided evidence for both quantitative and qualitative changes in complexes that specifically bound to a nucleotide motif (TTTGTAGGG) upstream of the transcriptional start site (Shiels et al., 2000). Thus, an electrophoretic mobility-shift assay (EMSA) found that factors derived from parasite enriched nuclear extracts of the macroschizont stage formed two complexes and that both complexes were detected at greater levels during merogony. This confirmed that factor(s) with the potential to regulate expression of Tams1 are present in preceding stage of the life cycle and increase during differentiation. A third complex, not detected with macroschizont extracts, was associated with the transition to high-level merozoite gene expression and production. Mutagenesis of TTTGTAGGG core motif confirmed its importance for the specific binding and revealed that flanking nucleotides may act to stabilise motif-factor complex (Shiels et al., 2000). Support for this model has recently been generated for *Plasmodium*. Thus in both *P. berghei* and *P. falciparum* lines that have lost the ability to generate gametocytes production, mutations or reduced expression levels were discovered in a putative transcription factor. It was subsequently shown that this transcription factor

(AP2-G) is required for commitment and is likely to operate via a positive feedback auto-regulatory loop (Kafsack et al., 2014; Sinha et al., 2014).

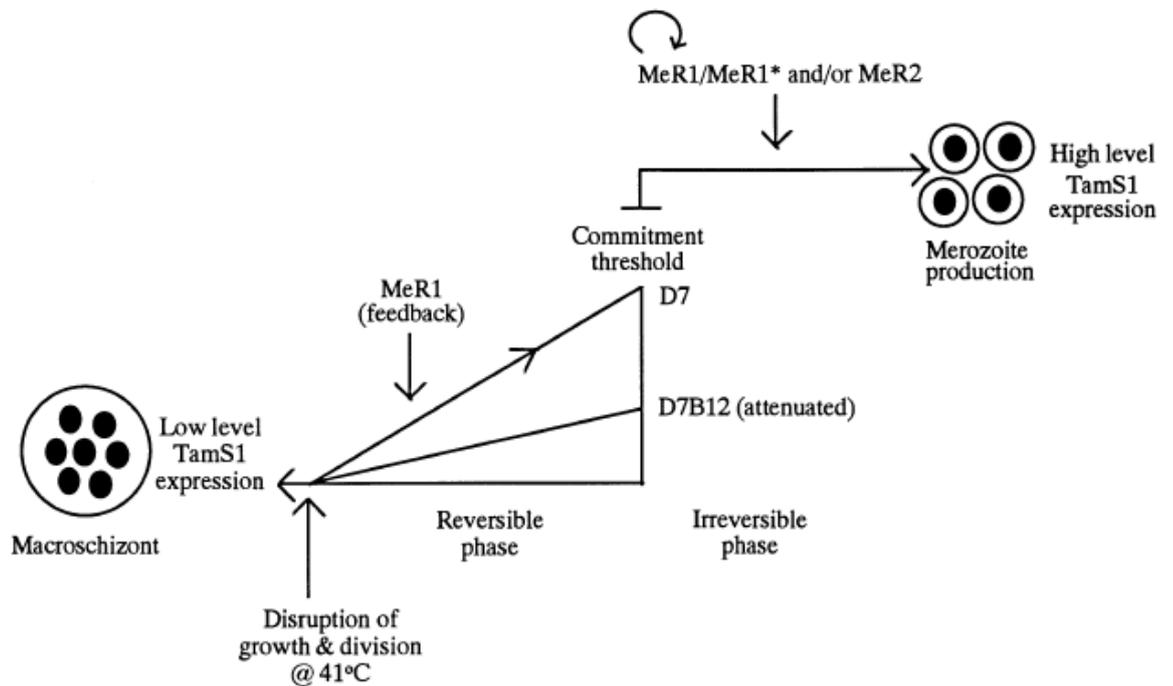


Figure 1.5 Theoretical molecular model for stage differentiation from macroschizont to merozoite of *T. annulata in vitro*. TamS1 - the major merozoite surface antigen gene, MeR - a regulator of merozoite gene expression (Shiels et al., 1998).

1.7 Regulation of gene expression in apicomplexan parasites

Progression from one parasite life-cycle stage to the next, with both entry and exit from the cell cycle, requires a strict programme of expression events, with co-expression of subsets of genes linked to different stages. Differential gene expression is required to facilitate different aspects of parasite stage development, such as biosynthesis and organisation of structural components for multiple morphological forms, the generation of parasite-encoded molecules that interact with or evade the host immune system (Hakimi and Deitsch, 2007) and the establishment of different modes of energy production and motility.

For certain parasites such as *Trypanosoma* or *Leishmania sp.*, gene expression mechanisms are unique with multi-cistronic transcription and regulation of mRNA stability playing a major role (Johnson et al., 1987; Myler et al., 2011). However, for Apicomplexa including *T. annulata*, differential regulation of gene expression can occur at the transcriptional level in a monocistronic manner (Roos et al., 1994; Crabb and Cowman, 1996; Horrocks et al., 1998).

In both prokaryotes and eukaryotes there are three major components of the transcriptional machinery: the RNA polymerase complex and associated protein complexes responsible for initiation and elongation of the transcript; the general transcription factors, that bind to the core promoter of single-stranded DNA gene templates and are required for the baseline expression of any gene; and the specific transcription factors (TFs) that bind to specific DNA sequences upstream of the coding region, thereby controlling the transcription of genetic information from DNA to mRNA (Lodish et al., 1999). These specific TFs act to either activate or repress the transcription of a gene. Examination of the *T. annulata* genome in comparison to those of other eukaryotes confirmed the presence of proteins involved in the majority of known nuclear processes, but a lack of conserved transcription factors binding to specific DNA sequences and regulatory motifs typically found in similar eukaryotic species was evident (Balaji et al., 2005; Pain et al., 2005), and this was unexpected as related apicomplexans and *T. annulata* have complex life-cycles. This paradox suggests that apicomplexan genomes encoded other undetected specific TFs that are unrelated or distantly related to previously identified DNA binding factors. Evolutionary ancestral relationship with photosynthetic organisms (Waller and McFadden, 2005) raises the possibility that Apicomplexa adapted their regulatory machinery from other species. Alternatively, it was proposed for Apicomplexa that stage-specific gene expression does not totally depend on a large number of specific transcription factors but involves, for example, post-transcriptional regulation by non-coding RNAs or gene-specific chromatin level modulation (Iyer et al., 2008). Epigenetic control of the mechanisms of gene expression appears to be conserved among the Apicomplexa (Gissot et al., 2009), although it is still under investigation and particularly for *Theileria* the role it plays in stage differentiation remains an open question.

The possibility that putative TFs had been overlooked in initial genome analysis was confirmed with the identification of a group of conserved putative Apetala2 (AP2) DNA-binding domains in a number of apicomplexan genomes. These were termed Apicomplexan Apetala2 (ApiAP2) transcription factors (Balaji et al., 2005). In addition, the discovery of cis-acting motifs that interact specifically with these transcription factors (Campbell et al., 2010) has provided a new foundation for the identification of transcription regulation mechanisms in other apicomplexan species. The ApiAP2 family was first identified in plants. In *Arabidopsis thaliana*, the AP2 family is the second largest class of transcription factors and contains approximately 145 members, which are involved in stress responses and the regulation of reproductive and vegetative organ development (Drews et al., 1991; Jofuku et al., 1994; Dietz and Vogel, 2010). In plants, AP2 proteins can function as either activators or repressors of transcription (De Silva et al., 2008). It was shown that the architecture of AP2 proteins is related to their function: single-domain AP2 genes are responsible for environmental stress responses in plants from thermo-tolerance to dehydration and the ethylene response, while proteins with two tandem AP2 domains, separated by a short, conserved linker sequence of 25 amino acids, are involved in regulating plant development (Balaji et al., 2005).

In apicomplexans, each ApiAP2 protein may contain from one to four AP2 domains, but most proteins have one, which is often the only globular domain in the entire protein. Double AP2 domain proteins are also found in a small family of bacterial proteins typified by DP2593 from *Desulfotalea psychrophila* (Balaji et al., 2005). All other AP2 domains found in bacteria (*Trichodesmum erythraeum*), viruses (*Enterobacteria phage RB 49* and *Bacteriophage Felix 01*) and mobile DNA elements contain a fusion of the AP2 domain with the EndoVII nuclease domain, at least two distinct members of the lambda integrase superfamily domain or one to two copies of a novel cysteine-rich domain containing five conserved cysteine residues (Magnani et al., 2004; Balaji et al., 2005).

Comparative studies across apicomplexan species and genera showed that there are:

- 35-43 copies of the ApiAP2 domain in 27 predicted proteins of *Plasmodium* species (*P. falciparum*, *Plasmodium chabaudi*, *Plasmodium yoelii* and *Plasmodium berghei*) (Painter et al., 2010);
- 24 copies of the ApiAP2 domain in *T. annulata*, corresponding to 22 predicted proteins (Balaji et al., 2005);
- 25-30 copies of the ApiAP2 domain in *C. parvum* and *C. hominis* (Balaji et al., 2005);
- 68 ApiAP2 proteins with one or more AP2 domains in *Toxoplasma* (Behnke et al., 2010).

The number of copies of the ApiAP2 domain appears to be positively correlated with genome size among apicomplexan species: *Toxoplasma* with a genome size approximately 80 Mb (Kissinger et al., 2003) has the highest number of ApiAP2 domains; *P. falciparum* and *C. parvum* with genomes of 22.8 Mb (Lau, 2009) and approximately 9 Mb (Puiu et al., 2004) respectively show a lower number of ApiAP2 proteins, while *T. annulata* with a genome of 8.3 Mb (Lau, 2009) has 24 copies.

Several studies have characterised binding specificities of *P. falciparum* ApiAP2 TFs (De Silva et al., 2008; Campbell et al., 2010; Flueck et al., 2010; Lindner et al., 2010, Sinha et al., 2014) and some of the ApiAP2 proteins have been confirmed to act as essential regulators of the parasite life-cycle (Yuda et al., 2009; Yuda et al., 2010, Sinha et al., 2014). DNA target motifs for ApiAP2 TFs have also been described for *T. gondii* (Behnke et al., 2010; Radke et al., 2013) and *C. parvum* (Oberstaller et al., 2013). Despite extensive sequence divergence between ApiAP2 proteins across distantly related apicomplexan species, the DNA-binding specificities of orthologous pairs of ApiAP2 domains are fundamentally conserved, although their downstream targets are not (De Silva et al., 2008; Campbell et al., 2010). The identified ApiAP2 domains of *C. parvum* and its *Plasmodium* orthologues have very similar DNA-binding specificities which confirms that the apicomplexan parasites have conserved not only the ApiAP2 DNA binding domain architecture, but also target sequence specificity (De Silva et al., 2008), which would suggest similar transcriptional regulation in *Theileria*

species. Moreover, analysis of the expression patterns of ApiAP2 genes in *P. falciparum* and *T. gondii* confirmed that these transcription factors are specifically expressed in different major developmental stages (Iyer et al., 2008; Painter et al., 2011; Walker et al., 2013; Radke et al., 2013) and are implicated in regulation of developmental transition. Therefore, it is likely that they may play an essential role in transcriptional regulation of *Theileria* stage-specific genes. An expression cascade of ApiAP2 proteins has also been observed in the IDC in *P. falciparum* (Campbell et al., 2010) and the tachyzoite cell cycle in *T. gondii* (Behnke et al., 2010). In addition to regulation of expression of their target genes, *P. falciparum* and *Toxoplasma* ApiAP2 proteins have been predicted to form a regulatory interaction network among themselves, regulating their own and other ApiAP2 expression by binding to target motifs in upstream regions (Campbell et al., 2010; Bougdour et al., 2010).

1.8 Aims and objectives

How parasites progress to the next stage of the life-cycle is an important area of research. The main aim of this project is to elucidate the molecular mechanisms underlying *T. annulata* stage-specific gene regulation. Evidence from the literature indicates that the differentiation process in *Theileria* involves binding of specific transcription factors to DNA motifs located upstream of stage-regulated target genes. Not only would validation of the Shiels molecular model of stage differentiation be important from a biological point of view, it would be potentially relevant to the characterisation of specific molecular targets for disruption of parasite transmission and thus may inform on the design of novel therapeutics or vaccines. Studies testing the validity of the stochastic differentiation model in *T. annulata* will require employing a combination of molecular and transcriptomic approaches in order to investigate putative transcription factors involved over the period of the differentiation event. Identification of TF target motifs and establishing binding specificities of TFs to these motifs would prove beneficial in determining whether stage-specific auto-regulation was a feature of gene expression control in *Theileria*, as in other eukaryotic systems (Serfling, 1989).

The ApiA2 proteins are excellent candidates for regulators of precisely timed and co-ordinated gene expression events associated with multi-stage differentiation steps in *T. annulata*. Investigating the relationship between different apicomplexan ApiAP2 TFs, and their gene targets, may also be important for establishing common regulatory processes across the Apicomplexa. Little is known of the transcription factors involved in regulation of the *Theileria* life-cycle, however definition of cis-acting DNA motifs associated with stage regulated genes or gene families would allow identification of target sites of putative regulatory factors. It may also provide greater understanding, based on target gene function, of events that are required for stage differentiation to be completed; rhoptry production in merozoites, for example. A better understanding of factors and target genes that operate to control stage differentiation is essential for rationale design of strategies aimed at disrupting stage development to inhibit establishment and transmission of tropical theileriosis in cattle.

The major objectives of the work presented in this thesis were as follows:

- I. Establishment, characterisation and validation of a subset of *T. annulata* genes that show significant differential expression during differentiation events in the bovine host.
- II. Bioinformatic identification and characterisation of novel DNA motifs found in the upstream regions of subsets of stage-regulated genes and investigation of their potential role in gene expression control.
- III. Identification and characterisation of ApiAP2 target motifs in upstream regions of *T. annulata* stage-regulated genes, previously found to act as ApiAP2 DNA binding sites in other Apicomplexa.
- IV. Characterisation of a selected panel of any differentially regulated *T. annulata* genes encoding ApiAP2 domains and generation of recombinant proteins.
- V. Comparison of expression levels among *T. annulata* genes encoding ApiAP2 domains between differentiation competent and attenuated cell lines.

- VI. Validation of binding specificities of *Theileria* ApiAP2 factors up-regulated from the macroschizont to the merozoite stage.
- VII. Validation of the binding potential of putative DNA binding motifs using parasite nuclear extracts from macroschizont infected cells and cells undergoing merogony.
- VIII. Characterisation of a potential ApiAP2 regulation network for merogony in *T. annulata*.

Chapter 2

Investigation of *Theileria annulata* gene expression during stage differentiation from macroschizont to merozoite

2 Investigation of *Theileria annulata* gene expression during stage differentiation from macroschizont to merozoite

2.1 Introduction

Stage differentiation is a fundamental aspect of apicomplexan biology, however, there is still very little known about the regulation of gene expression that facilitates this process. Characterisation of differentiation events of major apicomplexan parasites has shown that they share at least two differentiation steps, leading to postulation that events which control this process are conserved across the phylum (Shiels et al., 1998). *Theileria annulata* has a complex life cycle, with distinct stages occurring within the bovine host and tick vector, and is likely to share mechanisms that control stage differentiation steps with other *Apicomplexa* species. In this study, I focused on the bovine stages of the life cycle where multinuclear macroschizonts are formed within infected myeloid leucocytes and proliferate together causing transformation of the host cell. In vivo, proliferation of the infected cell slows down after a period of around 5 to 7 days and macroschizonts differentiate into uni-nucleated merozoites (Jarrett et al., 1969). Following destruction of infected leukocytes, released merozoites invade erythrocytes and form piroplasms. Piroplasm infected erythrocytes are then ingested by a feeding tick (Melhorn and Schein, 1984). In vitro, progression of the life cycle from macroschizont to merozoite stage can be induced by elevation of culture temperature to 41°C (pyrexia) (Shiels et al., 1992).

Studies performed by Shiels et al. (1992) demonstrated that differentiation to the merozoite is associated with a reduction in host cell proliferation together with an increase in parasite size. Furthermore, it was subsequently shown that the differentiation process is always asynchronous, but the longer a culture is incubated at 41°C, the greater is the probability that macroschizonts will differentiate (Shiels et al., 1992). Moreover, it was established that different cloned cell lines that appear to be isogenic, display markedly different abilities to undergo differentiation to the merozoite and that early passage lines can undergo differentiation at 37°C. These parameters indicate that differentiation from the macroschizont to the merozoite stage is a stochastic process, i.e. occurs with a random probability that can be altered by culture conditions.

Moreover, it was proven that in vivo production of *Eimeria* gametocytes is an asynchronous process (McDonald and Rose, 1987). It has also been found that asynchronous differentiation to the gametocytes in *Plasmodium* can be modulated by extracellular conditions (Carter et al., 1979). These studies suggest that stochastic differentiation is a common characteristic of many Apicomplexa.

To understand how the differentiation process is regulated in *T. annulata*, investigation of its stochastic nature was performed using drugs that inhibit DNA or protein synthesis. Cell cultures incubated with oxytetracycline (inhibitor of mitochondrial/cytoplasmic protein synthesis) increased the time period necessary to reach the commitment threshold, while aphidicolin (inhibitor of DNA synthesis) reduced it (Shiels et al., 1997). Based on these results, Shiels et al. (1999) proposed a stage differentiation model operating on the basis of factors that regulate merozoite gene expression reaching a concentration threshold which, once reached, commits the parasite to irreversible up-regulation of merozoite gene expression (see Chapter 1). The model also proposed that the probability of a differentiating event occurring would be dependent on the ratio of a protein regulator relative to the DNA target motif to which it binds. While recent work has supported the stochastic model of differentiation (Schmuckli-Maurer et al., 2008), only limited information on motifs and factors that could operate to regulate stage-specific gene expression in *Theileria* has been obtained (Shiels et al., 2000). However, demonstration that the probability of a differentiation event occurring can be altered highlights its potential as a target of control.

Bioinformatic analysis of a number of apicomplexan genome sequences have shown that certain transcription factors exhibiting DNA binding domains are absent in apicomplexan relative to other eukaryotes (see Chapter 1). The paucity of specific TFs in *T. annulata* and other apicomplexans was surprising, because they show a complex developmental life cycle within their hosts and also a reproducible pattern of differential gene expression (Balaji et al., 2005). In this regard, discovery of the ApiAP2 DNA-binding protein family by Balaji et al. (2005), and comparative analysis of apicomplexans such as *Plasmodium*, *Theileria*, and *Cryptosporidium* (Iyer et al., 2008; Campbell et al., 2010) has

shown that in the *T. annulata* genome there are 26 copies of the AP2 domain (22 predicted proteins) (Balaji et al., 2005) (Table 2.1). Moreover, relatively few additional hypothetical transcription factors have been identified in the genome sequences of *T. annulata* and *T. parva*, based on orthology with motifs from other eukaryotes (Table 2.2) (Iyer et al., 2008), highlighting ApiAP2 proteins as potentially key candidates for transcription regulators in *Theileria* parasites.

Domain(s) present in the protein	Gene name	Protein length (amino acid residues)	Species	Genbank protein description
ap2	TA09965	400	<i>T. annulata</i>	hypothetical protein
ap2	TA08375	409	<i>T. annulata</i>	hypothetical protein, conserved
ap2	TA07100	425	<i>T. annulata</i>	hypothetical protein, conserved
ap2	TA16535	590	<i>T. annulata</i>	hypothetical protein, conserved
ap2	TA11665	784	<i>T. annulata</i>	hypothetical protein, conserved
ap2	TA10940	396	<i>T. annulata</i>	hypothetical protein, conserved
zz+myb+swir m+ap2	TA06995	1146	<i>T. annulata</i>	transcriptional adaptor (ADA2 homologue), putative
ap2+ap2+ap2	TA05055	751	<i>T. annulata</i>	hypothetical protein
ap2+ap2	TA20595	718	<i>T. annulata</i>	hypothetical protein
ap2	TA12015	284	<i>T. annulata</i>	hypothetical protein
ap2	TA02615	300	<i>T. annulata</i>	hypothetical protein
ap2	TA04435	370	<i>T. annulata</i>	hypothetical protein
ap2	TA11145	578	<i>T. annulata</i>	hypothetical protein
ap2	TA18095	370	<i>T. annulata</i>	clathrin adapter complex-related protein
ap2	TA19920	830	<i>T. annulata</i>	hypothetical protein
ap2	TA13515	604	<i>T. annulata</i>	hypothetical protein
ap2	TA17415	193	<i>T. annulata</i>	hypothetical protein
ap2	TA16105	288	<i>T. annulata</i>	hypothetical protein
ap2+ap2	TA07550	541	<i>T. annulata</i>	hypothetical protein
ap2	TA16485	553	<i>T. annulata</i>	hypothetical protein
ap2	TA13395	248	<i>T. annulata</i>	hypothetical protein
ap2	TA04145	1158	<i>T. annulata</i>	hypothetical protein

Table 2.1. List of hypothetical ApiAP2 proteins in *T. annulata* (Iyer et al., 2008).

Domain(s) present in protein	Gene name	Protein length (amino acid residues)	Species	Genbank protein description
cbf	TA17930	424	<i>T. annulata</i>	hypothetical protein
gata	TA10735	574	<i>T. annulata</i>	GATA-specific transcription factor, putative
rpa	TA13200	466	<i>T. annulata</i>	hypothetical protein
rpa	TA15495	625	<i>T. annulata</i>	replication factor-A protein 1
znf+znf+znf	TA17985	261	<i>T. annulata</i>	zinc finger protein, putative
znf	TA19845	605	<i>T. annulata</i>	hypothetical protein
myb	TA17065	707	<i>T. annulata</i>	Myb-like DNA binding protein (CDC5 homologue), putative
snf+myb	TA06490	1012	<i>T. annulata</i>	SWI/SNF family transcriptional activator protein, putative
lrr+myb+myb	TA08790	1545	<i>T. annulata</i>	hypothetical protein
myb	TA12995	615	<i>T. annulata</i>	hypothetical protein
zz+myb	TA19000	610	<i>T. annulata</i>	hypothetical protein
swirm+myb	TA17880	588	<i>T. annulata</i>	hypothetical protein
myb	TA15395	578	<i>T. annulata</i>	hypothetical protein
zz+myb+swirm+ap2	TA06995	1146	<i>T. annulata</i>	transcriptional adaptor (ADA2 homologue), putative
myb	TA06455	665	<i>T. annulata</i>	hypothetical protein

Table 2.2. Hypothetical transcription factors of *T. annulata* other than ApiAP2 (Iyer et al., 2008).

From the genome sequence and the work of Balaji et al. (2005), putative DNA binding factors that could operate to regulate gene expression during stage differentiation in *Theileria* had been identified. An important aim of this thesis was to investigate further this possibility. One way to do this would be to a) determine whether the ApiAP2 factors themselves show evidence of differential regulation as the parasite differentiates and b) determine whether motifs that ApiAP2 factors (or other TFs) may bind to the DNA motifs present in the upstream regions of genes that are up- or down-regulated during the stage differentiation event. Therefore, identification of sets of genes up- or down-regulated during differentiation, which may be potential targets for ApiAP2 TFs, was the main aim of this chapter and will provide the base for future investigation of DNA motifs specific in their upstream regions. The analysis will include TashAT and SVSP gene families of unknown functions, but predicted to

be involved in host cell transformation (Schmuckli-Mauer, 2009; Shiels et al., 2006).

Members of the largest subtelomeric gene family reported for *T. annulata* and *T. parva* encodes the variable secreted proteins (SVSPs) (Schmuckli-Mauer, 2009). The encoded proteins are predicted to be secreted into the host cell cytoplasm and are thought to either contribute to host cell transformation or evade the bovine immune response (Schmuckli-Mauer, 2009). Based on EST (expressed sequence tag) data, most of the SVSP genes are indicated as expressed in macroschizont stage (Weir et al., 2010), and it can be predicted that they are down-regulated as the parasite differentiates from the macroschizont and transformation is reversed. Identification of a large gene family showing differential regulation may aid in identification of shared regulatory motifs.

A second gene family associated with the macroschizont stage that may be a target for differential gene expression during differentiation is the TashAT/TpHN family. The TashAT family is located on chromosome 1 and consists of 17 members. Six of these genes encode proteins bearing AT-hook DNA binding motifs (Johnson et al., 1988) and putative transcriptional transactivation domains (Swan et al., 2003). Several have been demonstrated to translocate to the host cell nucleus and bind to DNA. An orthologous cluster of 20 genes (TpHN) is found in *T. parva* (Swan et al., 1999; Swan et al., 2001; Shiels et al., 2005). Northern blotting has shown that a number of TashAT cluster genes (TashHN, TashAT2, TashAT3) are down-regulated early during the differentiation process, between day 2 and day 4 in vitro, while others show reduction in mRNA levels later in the time course (Tash1 and SuAT1) as the cell becomes committed to merozoite production (Swan et al., 2001; Shiels et al., 2004). These studies indicate that for the majority of the TashAT cluster gene expression will be down-regulated as merogony occurs but that different members show distinct temporal regulation.

Families of up-regulated genes have not been identified but there is clear evidence that significant up-regulation of gene expression occurs during merogony. Thus, it has been shown that the TamS1 gene is up-regulated at the transcriptional level during merogony, and a motif upstream of the mapped transcriptional start site was identified based on its ability to bind parasite

nuclear factors (Shiels et al., 2000). Furthermore, detailed comparison of down-regulated Tash1 expression relative to the up-regulated Tams1 gene (Swan et al., 1999; Swan et al., 2001) provided evidence for a coordinated temporal switch in gene regulation during merogony. The complexity of this process was not unravelled but was thought (Swan et al., 2001) to involve a switch over in factors that control gene expression at both the transcriptional and post transcriptional level. A second up-regulated gene identified during merogony encodes a rhoptry protein, with evidence that it is controlled further along a regulatory cascade, relative to the Tams1 gene (Shiels et al., 1997).

Despite identification of up- and down-regulated genes further work was required to generate data sets large enough to perform bioinformatic screening. The main aim of this chapter was to identify genes that show significant elevated or decreased levels of expression as *T. annulata* differentiates from the macroschizont to merozoite/piroplasm stages using high-throughput analysis of gene expression profiling generated by microarray. This would provide a data set to search for both nucleotide motifs and putative factors that may bind to such motifs and function as regulators of gene expression during merogony. Identification of these types of regulatory elements is essential for further investigation of the stochastic model of differentiation to the merozoite in *Theileria* parasites.

2.2 Materials and methods

2.2.1 Microarray design

Microarrays are a powerful tool which can be utilised to monitor the expression level of many genes simultaneously and identify those genes which are differentially expressed under different experimental conditions. A whole-genome tiling microarray approach was used to investigate *T. annulata* gene expression during differentiation from the macroschizont to the merozoite stage and to compare these with gene expression for two additional life-cycle stages, the sporozoite and the piroplasm. The genome sequence of *T. annulata* (Ankara C9) (Pain et al., 2005) was utilised to design a custom parasite tiling microarray (Weir, unpublished) consisting of abutted 45-mer oligonucleotide probes representing both DNA strands on each of the four nuclear chromosomes and the mitochondrial genome. The array was designed for use on a 1024 x 768 resolution chip and comprises 392,778 probes in total, 95 % of which are targeted to the *T. annulata* genome. The remaining probes comprise bovine gene-targeted probes or control probes, including a set of over 15,000 oligonucleotides with random sequence and of mixed GC content. Three biological replicates were used per time-point.

Oligonucleotide probes were mapped to the coding sequence of the parasite genome using the BLAST-like Alignment Tool (BLAT)(Kent, 2002) (Figure 2.1) and this allowed the identification of gene-specific sense and anti-sense probe sets to cover genes expressed on both strands. Parasite gene expression levels were determined using \log_2 -transformed median intensity value based on these gene-specific probe sets and the data was normalised using the Robust Multi-array Average (RMA) method (Irizarry et al., 2003).

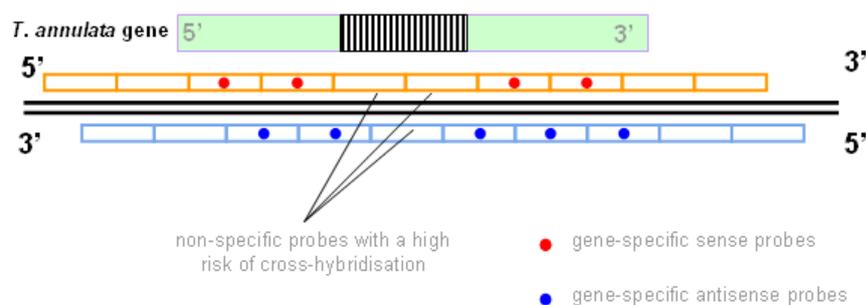


Figure 2.1. BLAT mapping of *T. annulata* genes
(www.theileria.org/ahdw/pictures/fig-5.gif;2013).

In order to determine whether a gene is expressed in a particular sample, the probe values for each gene were compared with the background values from non-specific, random probes with equivalent GC content. The relative levels of background hybridisation never exceeded a \log_2 intensity value of 10 and the highest background levels were found in the macroschizont Day 0 and Day 4 samples and the least background levels for piroplasm samples. Therefore, as a broad guideline, genes with a value of 10 or more were assumed to be expressed in a given parasite preparation. The proportion of parasite RNA in each preparation is estimated at 10 % in the macroschizont-infected cell, 50 % in infected cell cultures undergoing differentiation to the merozoite and 95 % or higher for purified piroplasm preparations.

2.2.1.1 Statistical analysis of microarray data

Gene expression profiling may be utilised to investigate the molecular mechanisms responsible for the regulation of parasite development, as it allows genome-wide analysis of the parasite transcriptome. DNASTAR ArrayStar3[®] software was used to perform hierarchical clustering and scatter plot analysis on \log_2 -transformed gene expression levels across the complete set of *T. annulata* genes. Hierarchical clustering is a method that groups data points into clusters by successively adding the data points into ever-growing groups using Euclidean distance metric which measures the difference in expression levels between the genes. The results of hierarchical clustering may be visualised as a heat map with the data clustered in two dimensions, i.e. by sample (vertically) and by gene (horizontally). Scatter plot analysis was initially performed to investigate the agreement between replicate samples and to broadly assess the relationship between macroschizont and merozoite/piroplasm gene expression values. A scatter plot is a type of diagram that can be used to explore the relationship between two quantitative variables, in this case gene expression values from different samples.

2.2.2 Rank Product analysis of differential gene expression

Rank Product (RP) analysis is a non-parametric statistical test that may be used to identify differentially expressed genes between conditions using limited sets of replicates (Breitling et al., 2004). Since Rank Products does not depend on an estimate of the gene-specific measurement of variance, it is particularly useful when only a small number of replicates are available. The normalised hybridisation intensity values obtained from the microarray (Weir, unpublished) were transformed into \log_2 expression values and subjected to RP analysis. The files generated from RP analysis of sporozoite to macroschizont, macroschizont to merozoite, merozoite to piroplasm, and piroplasm to sporozoite pair-wise gene expression comparisons were used in subsequent steps. The obtained RP score was used to rank all the *T. annulata* genes in the dataset according to the significance of their expression changes between stages and statistical confidence levels were assigned to each change in a form of false discovery rates (FDR) (Benjamini and Hochberg, 1995). Genes were identified as being up-regulated or down-regulated between different life-cycle stages based on two criteria, i.e. a fold change of greater than two and an associated FDR of less than 5%. Similar criteria have been previously applied in the analysis of microarray datasets (Benjamini and Hochberg, 1995; Jensen et al., 2006).

2.2.3 Cell lines and *in vitro* culture of *T. annulata*

Three cell lines were used in this study: D7, D7B12 and BL20. BL20 is an uninfected bovine lymphosarcoma cell line (Morzaria et al, 1982; Olobo and Black, 1989) frequently used in the analysis of *Theileria* macroschizont-infected cells as an uninfected control. D7 is a *T. annulata* (Ankara) infected cloned cell line that was subsequently re-cloned to produce D7B12 cell line (Shiels et al., 1994). Both clones appear to represent identical parasite genotypes (Shiels et al., 1994), however they display a marked difference in their ability to differentiate into merozoites when the cells are placed at 41 °C. D7 cells differentiate from macroschizont to merozoite with a high efficiency, while D7B12 cells remain in the macroschizont phase and do not differentiate to the merozoite stage (Shiels et al., 1994).

All cells used in this study were cultured in tissue culture flasks (25 cm² or 75 cm²) at 37 °C in the presence of 5 % CO₂. Cultures were maintained at 2 x 10⁵ cell/ml by feeding with fresh complete medium every two to three days. Each cell feeding cycle represented a 'passage' and corresponded to a 1 in 5 dilution of growing cells. Standard complete medium for *Theileria*-infected cells consisted of: 100 ml RPMI-1640 with added L-Glutamide and 25 mM HEPES (Gibco®); 25 ml Heat-inactivated foetal bovine serum (FBS; Sigma F9665); 200 µl Streptomycin/Penicillin (8 µg/ml); 300 µl Amphotericin B (0.6 µg/ml) and 1 ml 7 % sterile sodium bicarbonate.

2.2.3.1 Cryopreservation of cell lines

Long-term cryopreservation of the cell lines was undertaken using liquid nitrogen. 8 ml of cell culture was centrifuged at 1,000 rpm (182 x g) for 5 min at 4 °C and pellets were re-suspended in 3 ml complete medium containing 10 % dimethylsulphoxide (DMSO) and then split between two cryotubes. Cryotubes were wrapped in cotton wool and frozen in a pre-chilled polystyrene box at -80 °C overnight before being transferred to liquid nitrogen.

Cell recovery from liquid nitrogen was performed by quickly thawing cryopreserved tubes at 37 °C and immediately adding 8 ml pre-warmed complete medium, also at 37 °C. The cells were centrifuged at 1,000 rpm for 5 min at room temperature and then washed twice in 5 ml of pre-warmed fresh medium. Cell pellets were then re-suspended in 5 ml fresh complete medium and transferred to 25 cm² culture flasks. Overnight incubation was performed at 37 °C with 5 % CO₂. The next day a further 5 ml of complete medium was added and cells were passaged as normal.

2.2.3.2 *T. annulata* differentiation time-course

The *T. annulata* D7 cell line was used to perform a differentiation time-course experiment. The material generated from this experiment was used to validate the previous microarray data and for further analysis of gene expression changes and alterations in nuclear factor concentrations during merogony. D7B12 cells, which lack the capability to differentiate, were used as a control. Cells were cultured in standard complete medium (see 2.2.3) and induced to differentiate

by increasing the temperature from 37 °C to 41 °C. Cultures were established at 1.4×10^5 /ml and split to this concentration at Day 2 and Day 4. Since cell division slows down as differentiation progresses, cultures were diluted 1 in 2 at Day 6 or Day 7. Cells were harvested by centrifugation (4,000 rpm (3023 x g) for differentiating cultures) and total RNA isolated at Day 0, 4, 7 and 9. As determined in previous studies (Shiels et al. 1992): Day 0 represents the macroschizont stage; Day 4 is an intermediate, reversible time-point immediately prior to commitment for the majority of cells; Day 7 and Day 9 represent post-commitment time-points when merozoites are being generated by a significant component of the induced culture.

2.2.3.3 Giemsa staining

Morphological assessment of differentiating cells was undertaken by examining Giemsa stained samples at each culture time point. First, 3×10^4 cells/cm³ were deposited onto glass slides using a Shandon type 3 cytopspin at 1,500 rpm for 5 min. Slides were left to dry at 37 °C for 20 min, then fixed in ice-cold methanol for 30 min, dried again at 37 °C for 10 min and finally stained in 4 % Giemsa solution. After 30 min, slides were washed in tap water and left to dry at 37 °C. The slides were viewed under oil at x100 using an Olympus BX60 microscope and images were acquired using a SPOT digital camera and SPOT™ Advanced Image software Version Mac: 4.6.1.26 (Diagnostic Instruments, Inc).

2.2.4 Total RNA isolation and processing

RNA from the macroschizont stage and from differentiating macroschizonts was extracted directly from parasitised bovine cell cultures to provide template material for RT-PCR. Total RNA was isolated from D7 cells with TRI Reagent (Sigma; T9424) according to the manufacturer's instructions. Cells were scraped to detach them from the flask walls. Cell pellets were re-suspended in TRI Reagent (1 ml of TRI Reagent per 5×10^6 cells) and transferred to RNase-free 14 ml polypropylene Sarstedt tubes. Insoluble debris was removed by centrifugation at 12,000 g for 10 min at 4 °C. The supernatant was transferred to fresh tubes and allowed to stand at room temperature for 5 min to facilitate complete dissociation of nucleoprotein complexes. 0.2 ml chloroform per 1ml

TRI Reagent was added and mixed with the lysate by vigorous shaking for 15 sec and left at room temperature for 15 min. Following centrifugation at 12,000 g for 15 min at 4 °C, the aqueous phase at the top containing RNA was transferred into fresh tube. RNA was precipitated by adding 0.5 ml isopropanol per 1 ml of TRI Reagent at room temperature for 10 min. Precipitated RNA was pelleted at 12,000 g for 10 min at 4 °C, washed once with ethanol (1 ml 75 % ethanol per 1 ml TRI Reagent), briefly dried and then dissolved in 100 µl diethylpyrocarbonate (DEPC) treated water.

RNase contamination was minimised by use of DNase/RNase free plasticware and pipette tips. To eliminate possible contamination with genomic DNA, RNA samples were treated with DNase I (Qiagen; 79254) according to the manufacturer's protocol. DNase digestion was performed on approximately 100 µg RNA per tube with the use of RDD buffer (Qiagen, 79254) directly before column purification on a silica-gel membrane-based RNeasy column (Qiagen; 74104). Treated RNA was diluted in 30 µl of RNase-free water and absorbance was measured using a NanoDrop™ Spectrophotometer at 260 nm and 280 nm: an A260/A280 ratio of two or greater was indicative of good quality RNA. Purified RNA was stored in aliquots at -80 °C.

The quality and integrity of the total RNA was assessed by gel electrophoresis (1.2 % agarose TBE gel, with ethidium bromide, 100 V, 1 h) with UV light visualisation of 28S and 18S ribosomal RNA subunits using a FluorChem 5500 transilluminator system and image capturing software (Alpha Innotec; San Leandro, CA USA).

2.2.5 Reverse-Transcription Polymerase Chain Reaction

To confirm the results of the microarray analysis, Reverse-Transcription Polymerase Chain Reaction (RT-PCR) was performed in a semi-quantitative and quantitative (qRT-PCR) manner. RT-PCR is a powerful method for detecting and quantifying gene expression. mRNA transcripts act as a template for the synthesis of complementary DNA (cDNA) which is, in turn, used as a template for PCR (Figure 2.2).

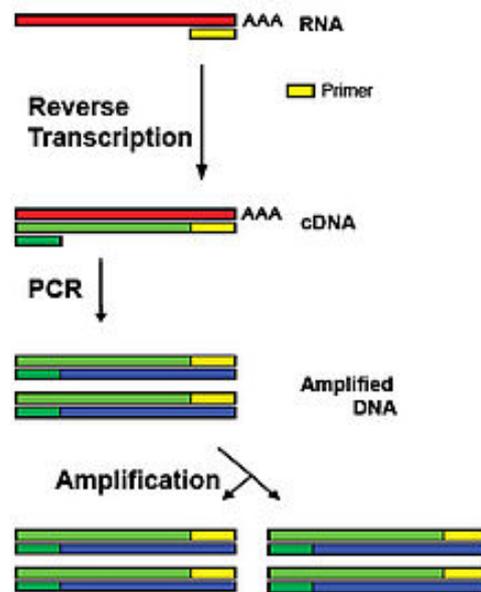


Figure 2.2. Reverse Transcription PCR

(http://upload.wikimedia.org/wikipedia/en/c/c2/Reverse_transcription_polymerase_chain_reaction.jpg).

In the present study, the microarray data for a number of differentially expressed genes identified from macroschizont to merozoite differentiating cultures and from comparisons of distinct life-cycle stages were validated by a one-step semi-quantitative RT-PCR approach where the entire reaction from cDNA synthesis to PCR amplification occurs in a single tube. Amplicons were then subjected to gel electrophoresis (see section 2.2.5.3).

2.2.5.1 Primer design and analysis

Primers for standard semi-quantitative RT-PCR were designed according to the following criteria: length 18-26 nucleotides; predicted amplicon size 220-350 bp; gene-specific in *T. annulata* and with minimal risk of amplification in the *Bos taurus* genome (Primer-BLAST; at least 5 total mismatches to unintended targets, including at least 2 within the last 5bp at 3' end). The melting temperature of the primer pairs was matched and was always between 50-58 °C.

Well designed primers for qRT-PCR are a prerequisite for a sensitive assay and the generation of specific amplicons. As SYBR Green I (see section 2.2.6) binds to any double-stranded DNA, it is important to reduce the risk of DNA

contamination by designing the primers that avoid forming dimers. Primers were designed to be between 18 and 30 nucleotides in length and with a GC content of between 40 % and 60 %. Amplicons were limited to 200 bp in size in order to obtain a high level of fluorescent dye integration without compromising the PCR efficiency. The melting temperature of the primer pairs was matched and was between 58-60 °C.

Primers for semi-quantitative and qRT-PCR were checked for uniqueness and specificity using the NCBI online Primer BLAST program (www.blast.ncbi.nlm.nih.gov/Blast.cgi). The Oligonucleotide Properties Calculator (www.basic.northwestern.edu/biotools/oligocalc.html) was used to avoid self-annealing sites and hairpin formation. Primers were synthesised by Eurofins MWG Operon (Ebersberg, Germany). Details of primer sequences, GenBank accession number, product length and annealing temperature for all gene-specific primers used in this study are listed in the Appendix (1.4). All primers were diluted to a stock concentration of 100 µM. The working concentration of all primers for semi-quantitative PCR was 5 µM. Initial PCR optimisation was performed to ascertain the ideal annealing temperature (T_a) and PCR efficiency. To maximise the sensitivity of the qPCR assay, the lowest concentration of primers possible was used.

2.2.5.2 Semi-quantitative RT-PCR reagents and cycling conditions

Semi-quantitative RT-PCR was performed in a final volume of 25 µl containing 20 ng of total RNA and using 35 cycles of PCR amplification. PCR reactions were performed using the SuperScript™ One-Step RT-PCR System protocol (www.tools.lifetechnologies.com/content/sfs/manuals/superscript_onestepRTPCR_man.pdf) in total volume of 25 µl with RT/Platinum® Taq DNA Polymerase (Invitrogen®; 12574-026) (0,5 µl), 2x Reaction Mix (a buffer containing 0.4 mM of each dNTP, 2.4mM MgSO₄) (12,5 µl), ddH₂O - 9µl, forward primer (5µM) (1 µl), reverse Primer (5µM) (1µl) and RNA (20ng/ul) (1µl). Thermal cycling conditions can be found the Appendix (1.2).

2.2.5.3 Gel electrophoresis and quantification of RT-PCR products

RT-PCR products were visualised on 1 % agarose gel (Sigma Aldrich) in 1 x TAE buffer and ethidium bromide. Gels were run in 1 x TAE buffer using a horizontal electrophoretic tank at 100 V for 1 hr. Visualisation and quantification of PCR product band intensity was performed relative to a control PCR based on the genes encoding the *T. annulata* heat shock (Hsp) 70 kDa protein (TA11610) and heat shock 90 kDa protein (TA07100). Microarray data suggests that these genes are constitutively expressed and do not show significant changes in expression from the macroschizont through merozoite to the piroplasm stage. Amplicon sizes were estimated using a 100 bp DNA ladder and gel images recorded using image analysis software Alphasampler FluorChem 5500 (Biosciences; Santa Clara, CA, USA).

2.2.6 Two-step quantitative Reverse-Transcription PCR

Real-time PCR product quantification by the fluorescence-kinetic detection method is popular technique for analysing gene expression on a small scale. One advantage of this method is that quantitative measurement of the PCR product is carried out during sequential amplification cycles and therefore no post-PCR manipulation is required. This methodology also enables comparison of the predicted melting temperature of the specific product to the observed results to determine if non-target products are present. Although one-step qRT-PCR minimises the possibility of cross-contamination, the two-step system is generally considered to be more sensitive and specific (Bustin, 2000), as the reaction conditions are optimised exclusively for both reactions. It also enables storing of cDNA samples and simultaneous quantification of several targets. Therefore, in order to fully validate the microarray results of selected differentially expressed genes, the SYBR Green qRT-PCR methodology was applied.

SYBR Green I is a cyanide intercalating dye which is commonly used as a fluorescent DNA-binding dye. It preferentially binds to double stranded DNA and

generates a fluorescent signal upon binding (Figure 2.3). In qRT-PCR, the DNA product accumulates during the amplification process and the fluorescent signal increases in proportion to the DNA concentration. The double stranded DNA-dye complex absorbs light of 497 nm and emits light at 520 nm, which makes it compatible with any real-time thermo-cycler. The initial copy number of cDNA targets can be quantified based on the threshold cycle (Ct), which is defined as the number of cycles required for the fluorescent signal to exceed the background signal. The threshold cycle is inversely proportional to the log of initial number of mRNA transcripts (Higuchi et al., 1993).

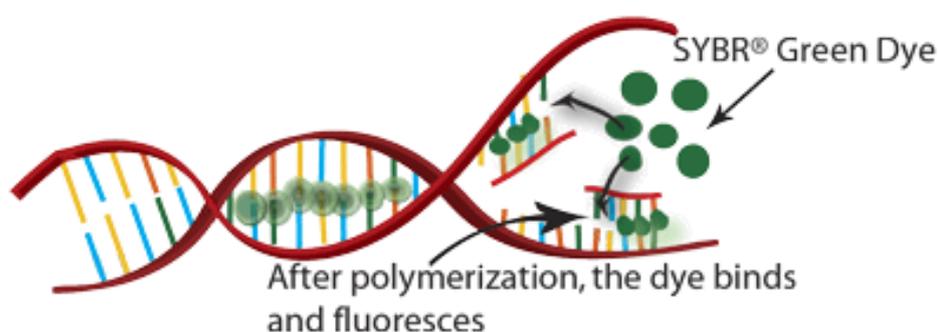


Figure 2.3. SYBR Green dye binding to the DNA.

(www.nfstc.org/pdi/Subject03/images/pdi_s03_m05_07_a.gif)

2.2.6.1 cDNA synthesis

cDNA synthesis was carried out in a total volume of 20 μ l using oligo(dT) primers provided with the AffinityScript Multi temperature cDNA Synthesis Kit (200436, Agilent) in a Techne thermo-cycler system (TC-512, Techne, UK) and following the supplier's protocol

(www.genomics.agilent.com/files/Manual/200436_B03.pdf).

First-strand cDNA synthesis reaction products were stored at -20°C . cDNA synthesis reaction components and thermal cycling conditions parameters can be found in Appendix (1.3).

2.2.6.2 Housekeeping genes

In order to normalise qRT-PCR data, it is necessary to select one or more reference genes with stable expression to correct artefactual variation caused by differences in cDNA quantity and quality that can affect the efficiency of the PCR reaction. In this study, the genes encoding the heat shock 70 kDa protein (TA11610) and heat shock 90 kDa protein (TA10720) were used for normalisation. These constitutively expressed genes did not show significant changes in expression levels during differentiation on the basis of both microarray and semi-quantitative RT-PCR data.

2.2.6.3 qRT-PCR reaction parameters

Quantitative RT-PCR was performed in a final volume of 25 μ l. Real-time fluorescence detection of PCR product was performed using Brilliant SYBR[®] Green QPCR Master Mix protocol (Stratagene, cat. no 600548) (www.chem-agilent.com/pdf/strata/600548.pdf). The qRT-PCR reaction details and thermal cycling parameters can be found in Appendix (1.5). SYBR[®] Green fluorescence detection was undertaken in 96-well Semi-skirted Flat Deck PCR plates (Thermo Fischer Scientific). All qRT-PCR data was captured and analysed by MxPro v4.10 software with the Mx3005P Real-Time PCR System (Agilent Technologies). Fluorescence was measured at every temperature increment of one degree Celcius. After 40 cycles of amplification, melting curve analysis was carried out to verify product specificity and determine the presence of primer-dimers and other non-target products. Technical replicates of each experimental time-point and no template controls were included in the PCR reaction for all samples.

2.2.6.4 qRT-PCR data analysis

Quantitative RT-PCR data was analysed using MxPro v4.10 software. The obtained data were converted using the comparative quantitation experiment option. The relative quantity values were normalised to the housekeeping genes Hsp70 and Hsp90 and fold-change was calculated relative to the calibrator

condition, Day 0 - macroschizont stage. This was performed using $-2^{-\Delta\Delta Ct}$ equation (Livak and Schmittgen, 2001), which determines the fold-change between the sample and control. This allowed graphs to be generated and data was presented as normalised mean values of \log_2 fold change \pm the standard error of the mean (SEM). Statistical analysis of qRT-PCR data representing D7 differentiation time points (Day 0, Day 4, Day 7 and Day 9) and the piroplasm stage (RNA preps of Day 14 of an *in vivo* infection, kindly donated by Dr Laetitia Lempereur, Glasgow) was performed using a One Way ANOVA test and post-hoc Dunnett test with Excel program (Microsoft) for two up-regulated from macroschizont to merozoite stage ApiAP2 genes (TA13515 and TA11145). Selection was based on microarray data (fold change ≥ 2 between macroschizont and merozoite stage) and literature indicating their potential role as orthologs of regulators of gene expression in other Apicomplexa species (Iyer et al., 2008; Campbell et al., 2010). Additional qRT-PCR reactions were performed for two other selected differentially expressed genes from macroschizont to merozoite stage (based on RP analysis) (TA10735 and TA15705) to confirm their down-regulation from macroschizont to merozoite stage. The difference between particular life stages was considered to be significant if the FDR associated with the gene was less than 0.05.

2.3 Results

2.3.1 Hierarchical clustering of *T. annulata* genes

The starting point for this project was a global analysis of microarray data to obtain data sets of genes that are up-regulated or down-regulated during differentiation from the macroschizont to merozoite/piroplasm stage of *T. annulata*. To detect similarities and differences across the microarray data generated for different life cycle stages and the *in vitro* differentiation time course, hierarchical clustering of \log_2 -transformed gene expression levels of all 3796 *T. annulata* genes was performed. The results are presented as a heat map in Figure 2.4. As expected from robust microarray data, three biological replicate samples per each time point cluster together and show strong similarity over the heat map profile. Each horizontal line represents an individual gene: green bands represent genes expressed at low levels, while black and red bands represent intermediate and highly expressed genes respectively. Based on this categorization it is clear that changes in expression values for particular genes occur across the X-axis of the heat map. Thus the detection of genes displaying expression profiles representing up- or down-regulation during differentiation from macroschizont to merozoite, and across the other stages represented by the array data, is evident. It can be seen also that the macroschizont 'Day 4' sample data clusters with 'Day 0' replicates, while the 'Day 7' replicates shows close similarity to merozoite 'Day 9' expression data. Clear differences between these two clusters are visible as Day 0 and 4 both represent macroschizont time points just prior to commitment; Day 7 and Day 9 on the other hand represent post-commitment time points when uninucleated merozoites are being generated. Sporozoite and piroplasm replicates clustered, as might be expected by order of life cycle, as they represent the beginning and the end of the tick phase, although major changes in expression levels between these two life stages are also evident. It can be concluded that a major changes in the control of gene expression, at the mRNA level, occur between Day 4 and Day 7 of differentiation to the merozoite *in vitro*. In addition, the heat map indicates a presence of clusters of a significant number of constitutively expressed genes with a high level of expression. They

correspond to predominantly black and red solid horizontal lines at the top of heat map.

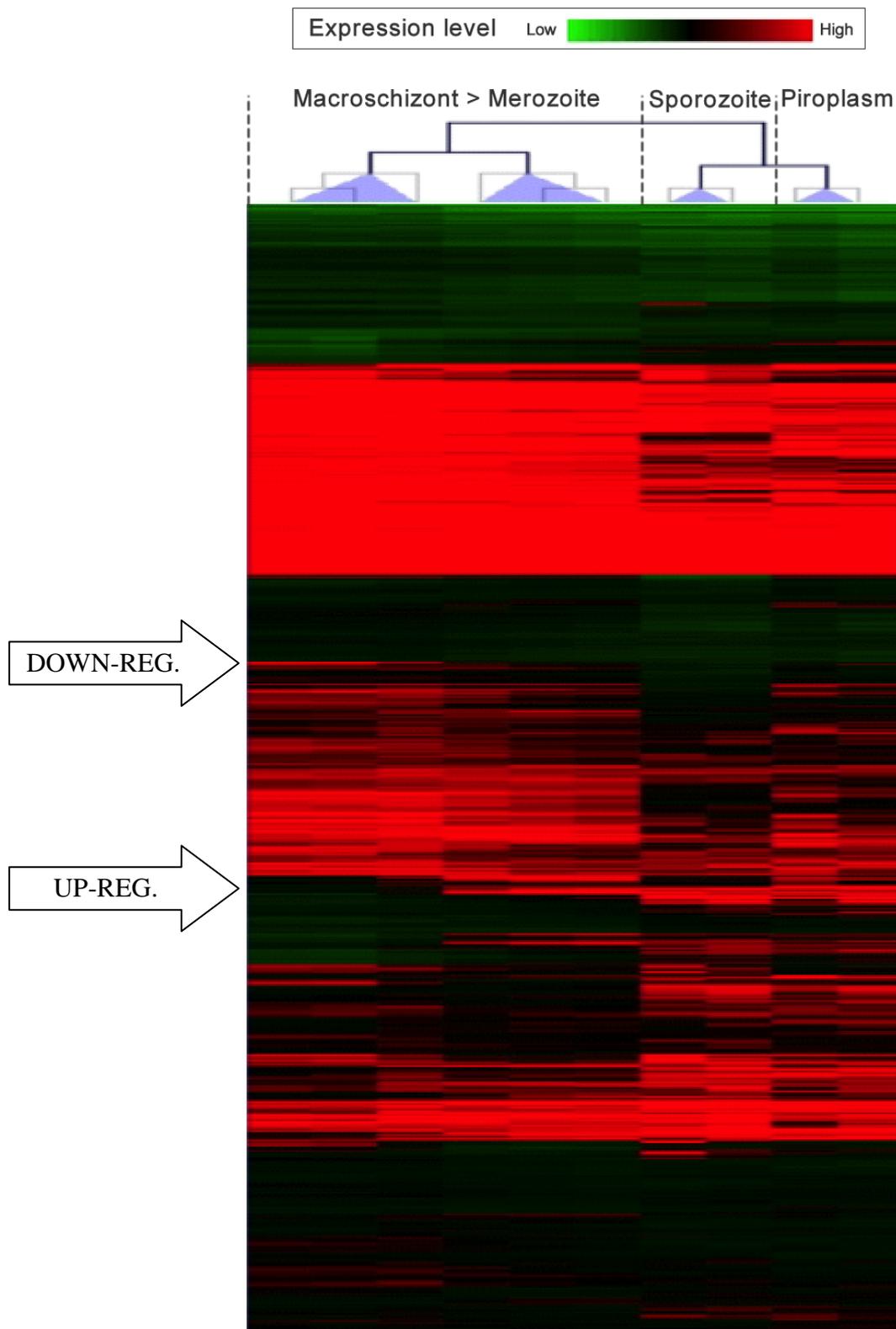


Figure 2.4. Visualisation of changes in gene expression level of 3796 genes identified by microarray analysis. Heat map of hierarchical clustering of all *T. annulata* genes shows differential expression across sporozoite, macroschizont, merozoite and piroplasm stages. Each horizontal line represents an individual gene. Green bands represent genes expressed at low levels, while black and red bands represent intermediate and highly expressed

genes respectively. White arrows indicate areas of differential expression - genes up-regulated from macroschizont to merozoite (UP) and down-regulated (DOWN).

2.3.2 Scatter plot analysis

To confirm statistically significant changes in gene expression and present the complete data set for different stage pair-wise comparisons, scatter plot analysis was conducted (RMA-normalized \log_2 expression values). The scatter plot represents a distribution of all 3796 genes, revealing up- and down-regulated genes between macroschizont and merozoite. Of the total number of genes used in the analysis, 587 genes showed an absolute change greater than 2-fold either up or down, 164 displayed a 4 fold or greater change and 43 an 8 fold or greater change (Figure 2.5). Each point along the scatter plot corresponds to a single gene. Blue dots represent genes expressed at low levels, while orange and red dots represent intermediate and highly expressed genes respectively. Three solid green lines are drawn diagonally across the scatter plot. The middle green line is the reference line and genes exhibiting equal expression levels lie along this line. The other two outer lines delineate genes with at least a two-fold change in expression level. Red line represents the best fit (trend) line.

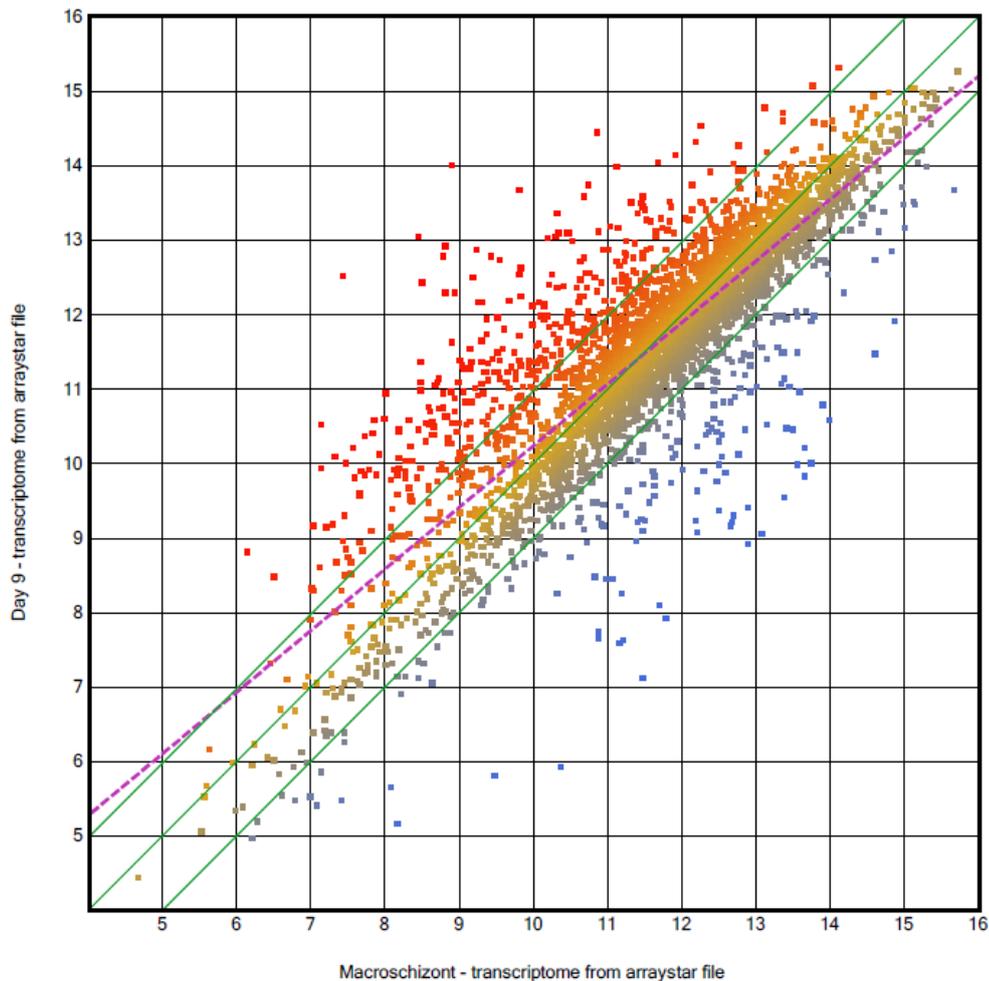


Figure 2.5. Scatter plot of macroshizont vs merozoite gene expression values. Blue dots represent genes expressed at low levels, while orange and red dots represent intermediate and highly expressed genes respectively. Two green outer lines delineate genes with at least a two-fold change in expression level between two life stages. Red dotted line - a trend line.

Scatter plot analysis was also performed to compare expression data across the macroshizont to the piroplasm stage to see if this would capture more differentially expressed genes i.e. those that may show a trend of up- or down-regulation from macroshizont to merozoite that only became significantly different when comparing macroshizont to piroplasm stage. As expected, an even larger number of up-regulated and down-regulated genes were identified (Figure 2.6). Out of 3796 analysed genes: 752 genes showed a 2 fold change or greater in expression value, 232 genes showed a 4 fold or greater and 74 showed a 8 fold or greater change. Black dots on the scatter plot represent the most up-regulated and down-regulated genes (8 fold change or greater). 48 of these genes are of parasite origin: bovine genes that were excluded from further analysis are marked by red circle. Bovine genes were placed on the array as a control and down-regulation of expression was expected, as macroshizont RNA

is derived from an infected leukocyte, where as piroplasm RNA is derived from parasites isolated from infected erythrocytes.

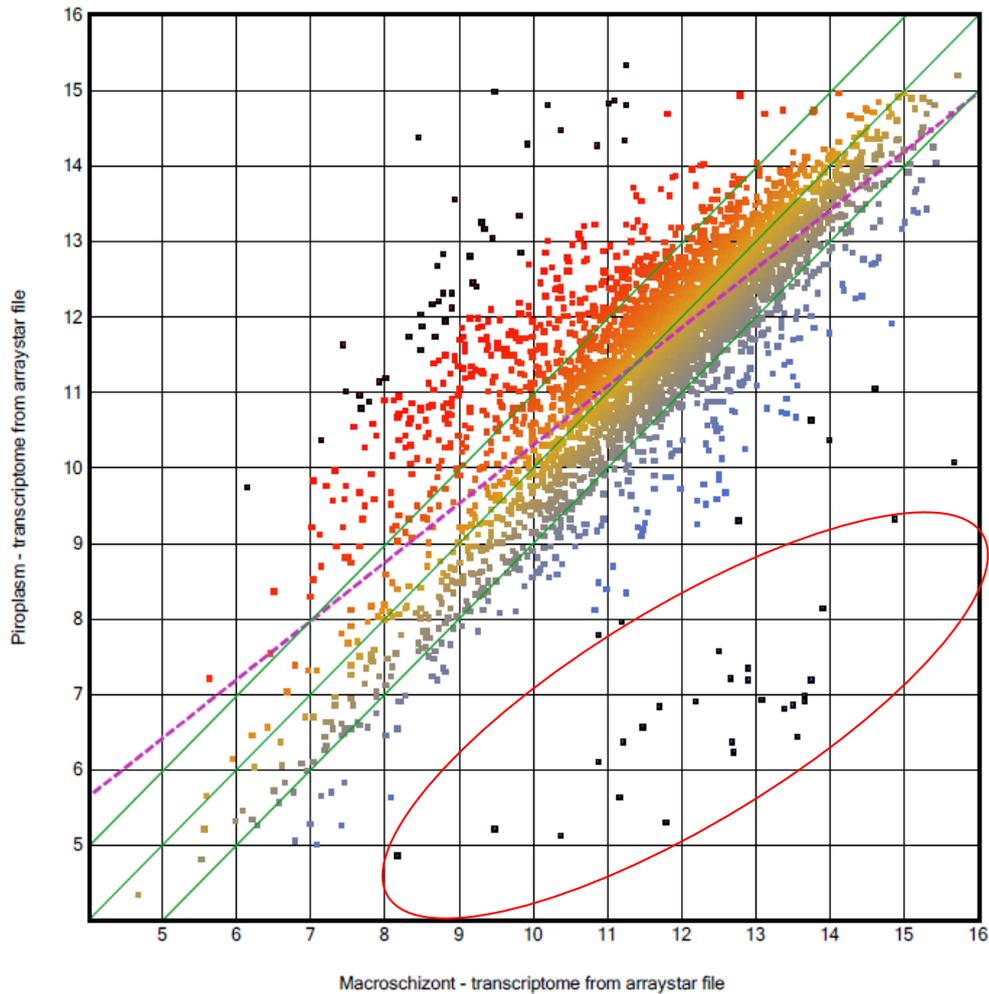


Figure 2.6. Scatter plot of macroschizont vs piroplasm gene expression values. Blue dots represent genes expressed at low levels, while orange and red dots represent intermediate and highly expressed genes, respectively. Black dots represent the most up-regulated and down-regulated genes with 8 fold change or greater. Two green outer lines delineate genes with at least a two-fold change in expression level between two life stages. Red dotted line - a trend line. Bovine genes marked by the red circle were used as a control for microarray assay.

2.3.3 Pair-wise comparison of genes differentially expressed between *T. annulata* life cycle stages

Following the general analysis of differential gene expression, pairwise comparison of normalized data (using RMA method) was performed using Rank Product (RP) analysis to support the initial identification of down- or up-regulated genes. This statistical method sets a cut-off for significance of differential gene expression values according to a false discovery rate (FDR). Applying a cut-off of $FDR \leq 0.05$ and fold change ≥ 2 (absolute), a summary table displaying the number of differentially expressed genes for each pair-wise comparison (across life-cycle stages) was generated (Table 2.3).

Between stage comparison	Fold change analysis, number of genes (≥ 2 -fold genes)		Rank Product, number of genes ($FDR \leq 5\%$)	
	Up	Down	Up	Down
Sporozoite to macroschizont	374	349	66	133
Macroschizont to merozoite	361	196	152	115
Merozoite to piroplasm	70	59	24	20
Piroplasms to sporozoite	239	182	57	35

Table 2.3. Summary data on number of differentially expressed genes obtained by pairwise comparison and RP analysis of microarray data sets of *T. annulata*.

The largest number of genes demonstrating a significant difference by RP was for the sporozoite to macroschizont (66 up-regulated genes and 133 down-regulated) and the macroschizont to merozoite (152 up-regulated and 115 down-regulated genes) pair-wise comparisons. The smallest number of genes was observed for the merozoite to piroplasm comparison (24 up-regulated and 20 down-regulated). Using an arbitrary 2-fold cut-off it was noticed that there is a greater number of genes up-regulated than down-regulated between *T. annulata* life-stages. Also there are much less significant changes in gene expression between merozoite to piroplasm stage in comparison to other life stages. This may have been expected, as previous studies indicated that the merozoite is more closely related to the piroplasm than the macroschizont (Shiels et al., 1992).

2.3.3.1 Identification of genes differentially expressed between the macroschizont to merozoite stage

Rank Product analysis identified the top 30 up- and down-regulated genes differentially expressed between the macroschizont and merozoite in vitro. These are listed in Tables 2.4 and 2.5 respectively. To extend the data set for subsequent analysis of expression profiles, combined with motif identification (see Chapter 3), the top 100 up- and down-regulated genes with an FDR<5% were obtained (see Appendix 2.1 and 2.2 for full list of genes obtained by RP). The list of 100 up-regulated genes mostly comprises genes encoding hypothetical proteins but also includes genes encoding rhoptry-associated proteins (TA05870 - the highest up-regulated gene from macroschizont to merozoite, TA05760 and TA05705) and three genes encoding hypothetical transcription factors, ApiAP2 domain proteins (TA13515, TA16485 and TA12015). Genes encoding a Map2 kinase (TA21080), a cysteine protease (TA04105, TA15660), myosin (TA20555), a phosphate transporter (TA13530), a ubiquitin-conjugating enzyme E2 (TA10690), a cyclin-dependent serine/threonine kinase-related protein (TA08470) and an aspartyl (acid) protease (TA17685) were also identified as being up-regulated, together with Tar/Tpr-related protein family members which have an unknown function (Weir et al., 2010).

	Gene ID	Annotation	FC (log ₂)	RP score (up)	EE (up)	FDR (up)
1	TA05870	rhoptry-associated protein, putative	5.12	3.13E-31	0	0
2	TA14665	hypothetical protein	5.13	1.41E-30	0	0
3	TA08360	hypothetical protein, conserved	4.64	1.69E-28	0	0
4	TA21080	Map2 kinase, putative	4.16	2.40E-26	0	0
5	TA05340	hypothetical protein, conserved	4.03	1.21E-25	0	0
6	TA16660	hypothetical protein, conserved	3.96	2.36E-25	0	0
7	TA05495	hypothetical protein	3.89	2.65E-25	0	0
8	TA13045	hypothetical protein, conserved	3.67	5.49E-24	0	0
9	TA13825	hypothetical protein	3.63	1.28E-23	0	0
10	TA07585	hypothetical protein	3.51	2.58E-23	0	0
11	TA19390	hypothetical protein, conserved	3.40	1.11E-22	0	0
12	TA18005	hypothetical protein	3.38	2.06E-22	0	0
13	TA19040	hypothetical protein, conserved	3.35	2.30E-22	0	0
14	TA19445	hypothetical protein, conserved	3.24	8.27E-22	0	0
15	TA11905	hypothetical protein	3.43	9.64E-22	0	0
16	TA20020	hypothetical protein, conserved	3.08	4.25E-21	0.02	0.001
17	TA14680	hypothetical protein	3.02	9.70E-21	0.03	0.002
18	TA21400	hypothetical protein	2.93	2.34E-20	0.04	0.002
19	TA17325	integral membrane protein, putative	2.95	2.41E-20	0.04	0.002
20	TA16375	hypothetical protein, conserved	2.94	3.47E-20	0.04	0.002
21	TA05760	rhoptry-associated protein, putative	2.91	3.70E-20	0.05	0.002
22	TA21395	hypothetical protein	2.95	4.08E-20	0.05	0.002
23	TA04105	cysteine proteinase, putative	2.90	5.75E-20	0.07	0.003
24	TA13515	hypothetical protein, conserved	2.88	8.26E-20	0.07	0.003
25	TA14955	hypothetical protein	2.85	1.43E-19	0.08	0.003
26	TA13215	hypothetical protein, conserved	2.79	1.43E-19	0.08	0.003
27	TA16485	hypothetical protein, conserved	2.78	2.13E-19	0.09	0.003
28	TA11455	hypothetical protein, conserved	2.76	2.77E-19	0.09	0.003
29	TA18855	Sfil-sub-telomeric fragment. related protein,	2.79	3.15E-19	0.1	0.003
30	TA16420	hypothetical protein	2.83	3.28E-19	0.1	0.003

Table 2.4. List of the thirty most up-regulated genes during differentiation from macroschizont to merozoite stage (ApiAP2 genes are highlighted).

Top of the list of 100 genes down-regulated during differentiation from the macroschizont to merozoite (Table 2.5) is the gene encoding a surface protein d precursor (TA19865). The list also included members of the two gene families encoding proteins predicted to be secreted into the host cell compartment and implicated in establishment of the macroschizont infected cell. Thus, members of the SVSP family and TashAT family (TA20095-TashAT2, TA03125, TA03120, TA03145 and TA03165) were found to be significantly down-regulated as the macroschizont differentiates to the merozoite. In addition, the gene encoding the macroschizont specific T cell antigen, Ta9 (TA15705) was found to be highly down-regulated. A gene encoding a predicted GATA-specific transcription factor (TA10735) was also found to be present among the most down-regulated macroschizont to merozoite genes.

	Gene ID	Annotation	FC (log ₂)	RP score (down)	EE (down)	FDR (down)
1	TA19865	surface protein d precursor	-3.38	7.04E-31	0	0
2	TA11410	Theileria-specific sub-telomeric protein, SVSP family. putative	-3.20	1.01E-28	0	0
3	TA11405	subtelomeric sfi-fragment-related protein family member. put.	-3.12	6.57E-28	0	0
4	TA15705	hypothetical protein (Ta9)	-3.10	2.17E-27	0	0
5	TA09805	Theileria-specific sub-telomeric protein, SVSP family	-2.93	7.76E-26	0	0
6	TA10505	hypothetical protein	-2.91	1.76E-25	0	0
7	TA09790	Theileria-specific sub-telomeric protein, SVSP family	-2.69	8.78E-24	0	0
8	TA18010	integral membrane protein, putative	-2.69	1.44E-23	0	0
9	TA09420	Theileria-specific sub-telomeric protein, SVSP family. putative	-2.61	2.04E-23	0	0
10	TA02480	hexose transporter (HT1 homologue), putative	-2.61	3.93E-23	0	0
11	TA05580	Theileria-specific sub-telomeric protein, SVSP family	-2.59	4.48E-23	0	0
12	TA15695	hypothetical protein	-2.64	4.76E-23	0	0
13	TA09810	Theileria-specific sub-telomeric protein, SVSP family	-2.54	6.70E-23	0	0
14	TA09435	Theileria-specific sub-telomeric protein, SVSP family. putative	-2.52	3.11E-22	0	0
15	TA15710	hypothetical protein	-2.50	4.99E-22	0	0
16	TA18895	conserved Theileria-specific sub-telomeric protein. SVSP family	-2.43	7.76E-22	0	0
17	TA09430	Theileria-specific sub-telomeric protein, SVSP family. putative	-2.40	1.95E-21	0.01	0.001
18	TA17555	Theileria-specific sub-telomeric protein, SVSP family	-2.40	2.15E-21	0.01	0.001
19	TA11940	hypothetical protein	-2.39	2.68E-21	0.01	0.001
20	TA09815	SfiI-subtelomeric fragment related protein family member. put.	-2.36	3.82E-21	0.02	0.001
21	TA09800	Theileria-specific sub-telomeric protein, SVSP family	-2.37	4.30E-21	0.02	0.001
22	TA10530	hypothetical protein	-2.35	5.18E-21	0.02	0.001
23	TA17545	Theileria-specific sub-telomeric protein, SVSP family	-2.25	4.52E-20	0.05	0.002
24	TA05575	Theileria-specific sub-telomeric protein, SVSP family	-2.24	5.32E-20	0.05	0.002
25	TA19005	conserved Theileria-specific sub-telomeric protein. SVSP family	-2.23	7.05E-20	0.07	0.004
26	TA17125	Theileria-specific sub-telomeric protein, SVSP family	-2.19	1.21E-19	0.08	0.003
27	TA09505	SfiI-subtelomeric fragment related protein family member. put.	-2.18	2.01E-19	0.09	0.003
28	TA18890	conserved Theileria-specific sub-telomeric protein. SVSP family	-2.17	2.72E-19	0.09	0.003
29	TA20095	Tashat2 protein	-2.13	6.55E-19	0.15	0.005
30	TA02905	hypothetical protein	-2.04	8.61E-19	0.16	0.005

Table 2.5. List of the thirty most down-regulated genes during differentiation from macroschizont to merozoite stage.

2.3.3.2 Identification of genes differentially expressed between the sporozoite to macroschizont stage

Rank Product analysis of sporozoite to macroschizont gene expression values identified mostly hypothetical proteins for down-regulation genes including the Sfil-subtelomeric fragment related proteins (TA07435, TA17595, TA17115, TA13020, TA16700), but also contained genes encoding a putative rhoptry protein (TA05870), a Map2 kinase (TA21080), Tpr-related protein family members (TA04790, TA03855, TA15445), an aspartyl (acid) protease (TA17685), a cysteine proteinase (TA04105), a cysteine repeat modular protein homologue (TA20782), an endonuclease (TA02660), a cyclin-dependent serine/threonine kinase (Cdk)-related protein (TA08470), a calmodulin-like domain protein kinase (TA16180), a thrombospondin-related protein (TA07755), an RNA polymerase II carboxyterminal domain (CTD) phosphatase (TA02640) and a potential AP2 domain transcription factor (TA12015).

Genes displaying evidence of up-regulated expression from sporozoite to macroschizont included: Tpr-related protein family members and hypothetical proteins, a surface protein d precursor (TA19865), chaperonin (HSP60) (TA07065), bifunctional dihydrofolate reductase/thymidilate synthase (TA08775), ATP synthase beta chain, mitochondrial precursor (TA20945), protein disulphide isomerase (TA04450), aminopeptidase n (TA20910), a bacterial histone-like protein (TA08715), a polymorphic antigen precursor-like protein (TA16685), elongation factor 1-gamma (TA08705) and Theileria-specific subtelomeric proteins from the SVSP family (TA09805, TA09790). The full list of genes of these two datasets is available in the Appendix (2.3 and 2.4).

2.3.3.3 Identification of genes differentially expressed between the merozoite to piroplasm stage

Rank Product analysis of merozoite to piroplasm gene expression values identified the most up-regulated genes as: a Sfil-subtelomeric fragment related protein (TA05525), two cysteine proteinase precursors tacP (TA03730, TA03750), a Tpr-related protein family member (TA15095), Sfil-subtelomeric fragment

related protein family members (TA17605, TA17500, TA09450), a leucine carboxyl methyltransferase (TA09205), TashAT family member (TA03115) and a putative AP2 domain transcription factor TA13515.

The data set of the most down-regulated genes included: phosphoenolpyruvate carboxykinase (TA20590), pepsinogen (TA02750, TA03860), a membrane protein family member (TA12045), a *Theileria*-specific integral membrane protein (TA17865), a rhoptry-associated protein (TA05760), a Tpr-related protein family member (TA04460), an ABC-transporter protein family member (TA17365) and an integral membrane protein (TA20325). The full list of merozoite to piroplasm stage regulated genes is available in the Appendix (2.5 and 2.6).

2.3.3.4 Identification of genes differentially expressed between the piroplasm to sporozoite stage

The list of most up-regulated genes identified by RP for the pairwise comparison of piroplasm vs sporozoite contains: Sfil-subtelomeric fragment related protein family members (TA09505, TA12195, TA09510, TA02735, TA12140, TA12275, TA16700, TA11400, TA11405), the membrane protein family member (TA12045), *Theileria*-specific sub-telomeric proteins of the SVSP family (TA17550, TA05540, TA05580, TA17555, TA05555, TA09435, TA16040, TA17540), the sporozoite surface antigen SPAG1 (TA03755), a polymorphic antigen precursor (TA17375), hexokinase 1 (TA19800), TashAT family member proteins (TA03150-Tash1e and TA03130-SuAT2), an ABC transporter (TA05455) and the putative GATA-specific transcription factor (TA10735), suggesting that this transcription factor might be important for the production of sporozoites in the tick or the expression of parasite genes following sporozoite invasion of the leukocyte. Genes down-regulated in expression between the piroplasm and sporozoite stages include Tpr-related protein family members, hypothetical proteins and also the putative AP2 domain transcription factor TA13515 suggesting that its expression peaks in the piroplasm or early phases of the parasite life cycle within the tick (gametocyte to kinete). The full list of piroplasm to sporozoite stage regulated genes is available in Appendix (2.7 and 2.8).

2.3.4 Microarray life cycle expression profiles of putative transcription factors of *T. annulata*

Scatter plot analysis of macroschizont vs merozoite and macroschizont vs piroplasm (Figure 2.7) gene expression values was screened to identify genes encoding potential transcription factors (TFs) (see Table 2.1 for list of putative ApiAP2 TFs identified for *T. annulata*). Fold change analysis of expression values for the macroschizont vs merozoite differentiation step identified four TF genes with an absolute fold change ≥ 2 (TA13515 - fold change of 7.36; TA16485 - 6.84; TA12015 - 4.01; TA11145 - 3.08). For the macroschizont vs piroplasm data set, up-regulation of six genes was identified: AP2 domain family members, TA13515 (absolute fold change (FC) of 25.46), TA12015 (FC, 5.58), TA16485 (FC, 4.28), TA11145 (FC, 2.65), TA10940 (FC, 2.49), TA17415 (FC, 2.02). Only three putative TF genes showed a statistically significant change in expression level that indicated reduced expression from macroschizont to piroplasm (≥ 2 fold change): TA10735 - the GATA transcription factor (fold change of 2.65 reduced in piroplasm), and two members of ApiAP2 family TA16536 (FC, 2.37) and TA13395 (FC, 2.24). No significant fold change values indicating a down-regulated TF from macroschizont to merozoite were observed.

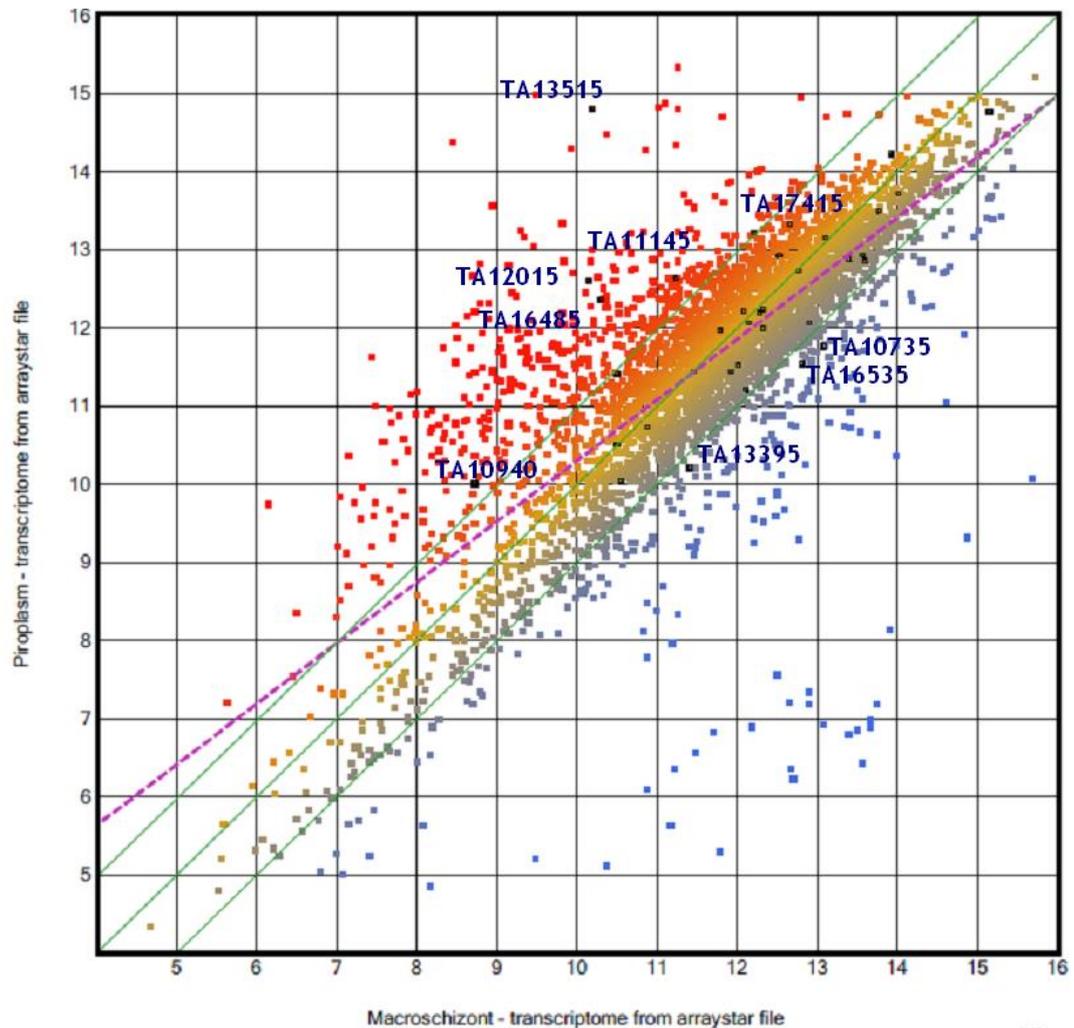


Figure 2.7. Scatter plot of macroschizont vs piroplasm gene expression \log_2 values with *T. annulata* potential transcription factors highlighted in black. Blue dots represents genes expressed at low levels, while orange and red dots represent intermediate and highly expressed genes respectively. Two green outer lines delineate genes with at least a two-fold change in expression level between two life stages. Red dotted line - a trend line.

To further understand the biological function of the ApiAP2 proteins in *Theileria annulata*, especially in the context of stage differentiation, the expression data obtained for the life cycle stages from sporozoite through to piroplasm was exploited. Twenty two *T. annulata* ApiAP2 genes were clustered based on their expression patterns using K-means clustering. At K=5, this gave rise to four major clusters, with each comprising 4-6 distinct ApiAP2 genes. In the cluster A four ApiAP2 genes (TA13515, TA16485, TA11145, TA12085) were noticeably up-regulated and showed the greatest level of modulation compared to the rest of the gene family (Figure 2.8.A). This suggests that these ApiAP2 factors may be functionally required to be up-regulated by the merozoite or piroplasm stages and could play a role in the regulation of these stage differentiation event. It

should also be noted the three of these genes in particular, showed a marked reduction in RNA levels in the macroschizont stage relative to all other analysed stages. During the in vitro time course TA13515, TA11145 and TA12015 genes shown an elevation to Day 4 followed by slight down-regulation or no change in expression levels between day 4 and 7 and then a further elevation to Day 9. TA16485 showed a delayed elevation compared to the other three manifest as a sharp increase of gene expression level at day 4 that continued to Day 9. Of these four ApiAP2s only TA13515 showed a significant increase in expression levels from merozoite to piroplasm stage, the others displaying a non significant slight decrease or increase.

In profile B three ApiAP2 genes - TA16535, TA13395 and TA06995 (see Figure 2.8.B) were found to have an expression profile that indicated higher expression levels associated with the macroschizont stage, which then declined during differentiation to the merozoite. In profiles C and D (Figure 2.8.C and D) no major change to the trend in expression of the ApiAP2 genes was observed during the macroschizont to merozoite differentiation step. For some of these genes (eg TA07100, TA07550 and TA04145) higher expression in the stages associated with the white blood cell was indicated, while for others (TA10940, TA05055 and TA08375) the reciprocal trend was detected. Excluding the TA10940 and TA16105 genes, all *T. annulata* ApiAP2 genes showed expression values of 10 or more and were predicted to be expressed by the parasite from macroschizont to merozoite stage at the RNA level. As the level for TA10940 and TA16105 was above 10 for the sporozoite, expression may be specific for this stage or stages present in the tick.

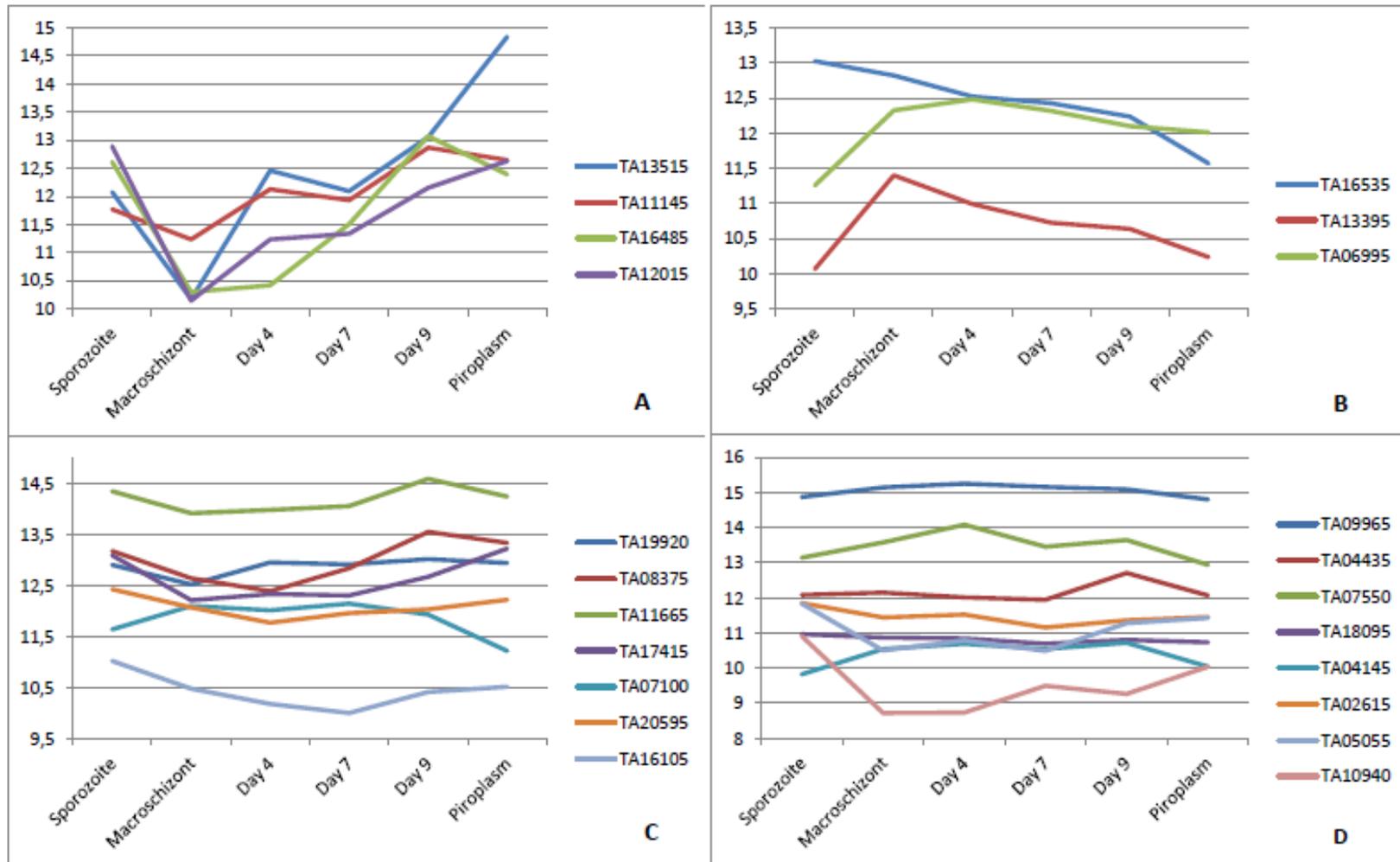


Figure 2.8. *T. annulata* ApiAP2 - temporal expression profile. A - genes up-regulated in macroschizont to merozoite stage; B - genes down-regulated in macroschizont to merozoite stage; C and D - genes at similar level in macroschizont to merozoite differentiation process. Y-axis denotes \log_2 expression values.

As shown in Figure 2.9 most of the other (non ApiAP2) predicted transcription factors in *T. annulata* did not show major variation to expression levels during macroschizont to merozoite differentiation (Figure 2.9.C). However several interesting changes to expression of individual genes were observed. For example significant down-regulation of TA10735 (GATA-specific transcription factor) gene from sporozoite to day 4 of differentiation was detected and this lower level of expression was then maintained through to the piroplasm (Figure 2.9.A). In contrast, TA06490 (SWI/SNF), TA06995 (ADA2 homologue) and TA17930 are elevated in stages associated with intracellular infection of the bovine leukocyte (Figure 2.9.B). Thus major changes to gene expression levels during the life cycle appear to operate for certain other predicted transcription factors.

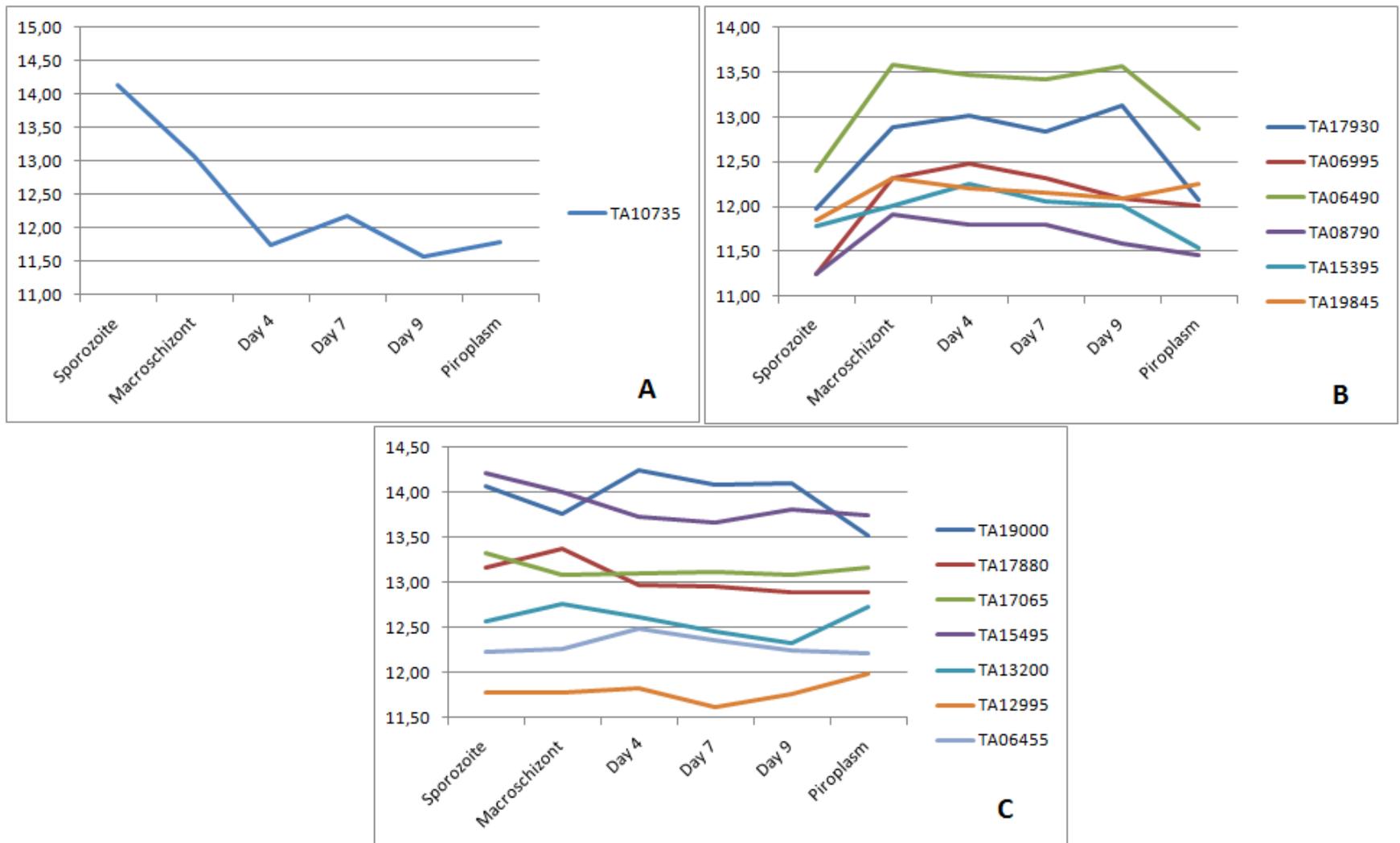


Figure 2.9. Temporal expression profile of non-ApiAP2 predicted transcription factors. A - genes down-regulated significantly in macroschizont to merozoite stage; B - genes up-regulated in macroschizont and merozoite stages; C - genes at similar level (no significant difference) in different life cycle stages. Y-axis denotes \log_2 expression values.

2.3.5 Differentiation from macroschizont to merozoite stage of *T. annulata* in vitro

To validate the microarray data set, a time-course of D7 cells differentiating from macroschizont to merozoite was required. This supplied a source of mRNA to perform semi-quantitative RT-PCR and qRT-PCR. RNA was isolated as described in section 2.2.3. Analysis of differentiation time-courses of *T. annulata* was performed by Giemsa staining of cytopins and microscopy for observation of morphological changes known to be associated with the differentiation process (Melhorn and Shein, 1984). Representative results are presented in Figure 2.10. Day 0 represents the macroschizont stage; Day 4 is an intermediate time point just prior to commitment for the majority of cells with enlarged multinuclear macroschizonts present within the infected leukocyte; Day 7 and Day 9 represent cultures post-commitment time points when uninucleated merozoites are being generated. At these points the parasite can virtually occupy the whole host cell cytoplasm, the parasite nuclei become enlarged and stain more densely, prior to the detection of small densely staining merozoite nuclei within the leukocyte cytoplasm (see Day 9). Destruction of infected leukocytes occurs and merozoites are released to the culture medium. As documented this process was always asynchronous with a greater proportion of the culture generating detectable merozoite nuclei from Day 6 onwards.

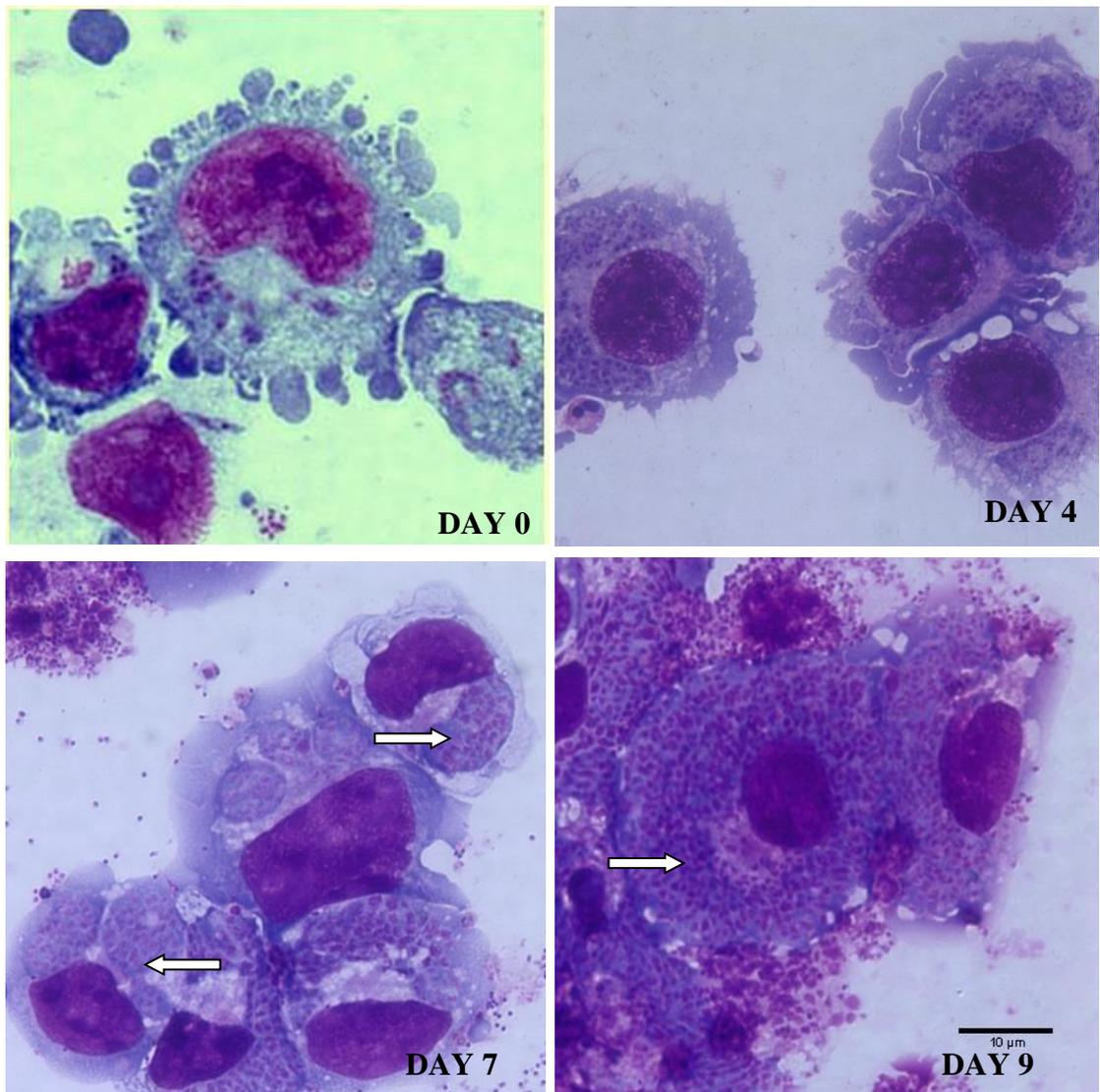


Figure 2.10. Giemsa stained lymphoblastoid cells infected with *T. annulata* following culture of the D7 cell line (x100) at 37°C (DAY 0), and at 41°C for 4 (DAY 4), 7 (DAY 7) and 9 days (DAY 9). Bar = 10 microns. Arrow heads denote merozoite nuclei.

2.3.6 Validation of differentially regulated genes by semi-quantitative RT-PCR

Six representative genes up-regulated from macroschizont to merozoite identified by Rank Product analysis were chosen for validation by semi-quantitative RT-PCR (Figure 2.11). This was performed using total RNA isolated from an in vitro time course of differentiation from macroschizont to merozoite (Table 2.6). The results confirmed the trend of up-regulation indicated by the array profiles for the selected genes. Some specific differences between time points were observed for certain genes when relating band intensity in comparison to the microarray results. This is likely to be caused by sampling/pipetting error and inaccuracies of the semi-quantitative RT-PCR, as well as a tendency of normalised microarray data to flatten out expression differences between conditions. However, it can be concluded that in general the semi-quantitative RT-PCR confirmed the microarray data indicating that the genes classified as significantly different are differentially expressed between macroschizont and merozoite.

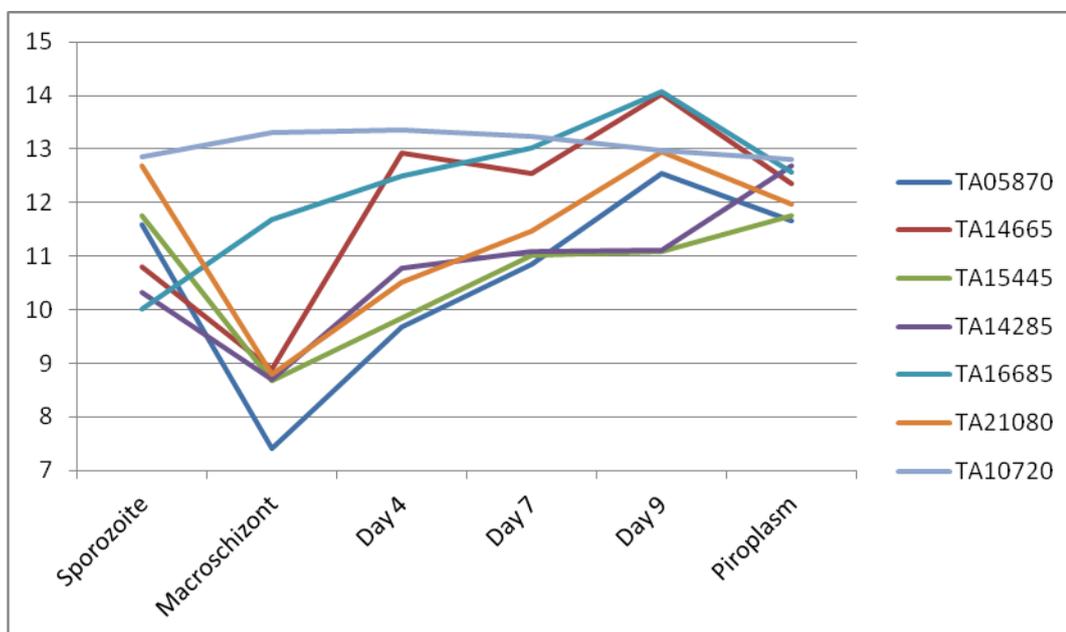


Figure 2.11. Temporal expression patterns of 6 representative of most up-regulated genes based on Rank Product analysis of microarray data. TA10720 gene (HSP90) included as a constitutive control. Y-axis denotes \log_2 expression values.

<u>GENE ID</u>	<u>Gene description</u>	Up-regulated genes – RT-PCR			
		Expression pattern on tRNA from macroschizont to merozoite			
		DAY 0	DAY 4	DAY 7	DAY 9
TA05870	rhoptry-associated protein				
TA14665	hypothetical protein				
TA15445	Tpr-related protein family member				
TA14285	Sfil-subtelomeric fragment related protein family member				
TA16685	polymorphic antigen precursor-like protein				
TA21080	Map2 kinase				
TA10720	HSP90 – constitutive control				

Table 2.6. Semi-quantitative RT-PCR of up-regulated genes from macroschizont to merozoite stage.

Additionally 6 representative genes of most down-regulated from macroschizont to merozoite stage based on Rank Product analysis were chosen for validation by semi-quantitative RT-PCR (Figure 2.12). As for the up-regulated set of genes the semi-quantitative RT-PCR results were consistent with a general trend of down-regulation of these selected genes between particular time points (Table 2.7).

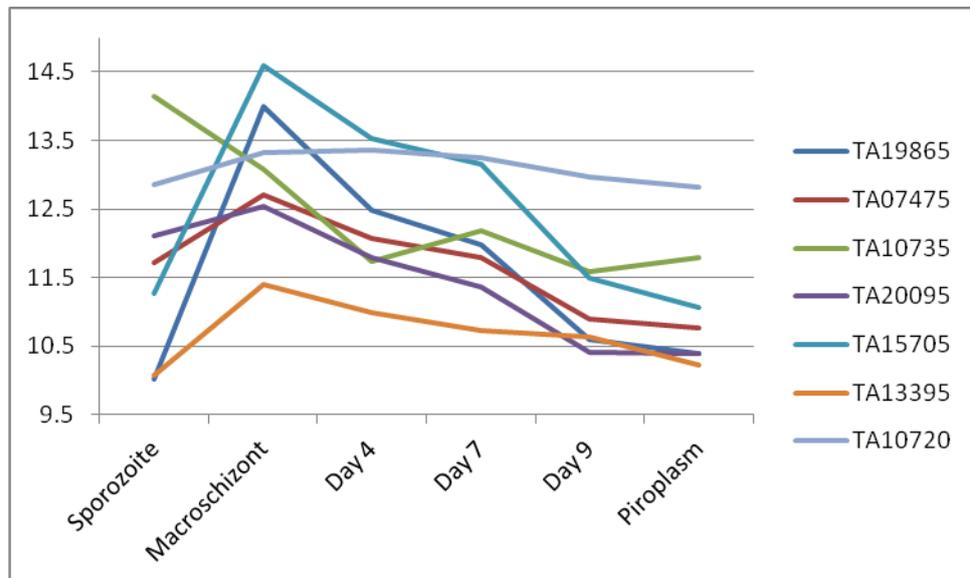


Figure 2.12. Temporal expression patterns of 6 representative most down-regulated genes based on Rank Product analysis of microarray data. TA10720 gene (HSP90) included as a constitutive control. Y-axis denotes log₂ expression values.

GENE ID	Gene description	Down-regulated genes - Expression pattern from macroschizont to merozoite DAY 0, DAY 4, DAY 7, DAY 9
TA19865	surface protein d precursor	
TA07475	Sfil-subtelomeric fragment related protein family member	
TA10735	GATA-specific transcription factor	
TA20095	TashAT2 protein	
TA15705	hypothetical protein (Ta9)	
TA13395	Hypothetical protein (ApiAP2)	
TA10720	HSP90 – constitutive control	

Table 2.7. Semi-quantitative RT-PCR of down-regulated genes on tRNA from macroschizont to merozoite stage in vitro culture.

2.3.7 Validation of differential expression patterns of selected potential transcription factors by RT-PCR

Based on analysis of expression changes between particular life-stages of *T. annulata* four up-regulated ApiAP2 genes (TA13515, TA12015, TA16485, TA11145) (Figure 2.13) and two down-regulated putative TFs (TA10735 (GATA) and TA13395 (ApiAP2)) (Figure 2.15) were chosen for analysis by semi-quantitative RT-PCR. As seen in Figures 2.14 and 2.16 the results confirmed up- and down-regulation of the TF genes during macroschizont to merozoite differentiation stage. Thus for TA13515, TA12015 and TA11145 both array and PCR data showed an initial elevation between Day 0 and 4 while for TA16485 elevation did not occur until after Day 4. Semi-quantitative RT-PCR was also performed for piroplasm stage for these genes and again mirrored the array data with up-regulation observed for TA13515 gene, and evidence of lower level expression of TA16485 compared to the Day 9/merozoite time point.

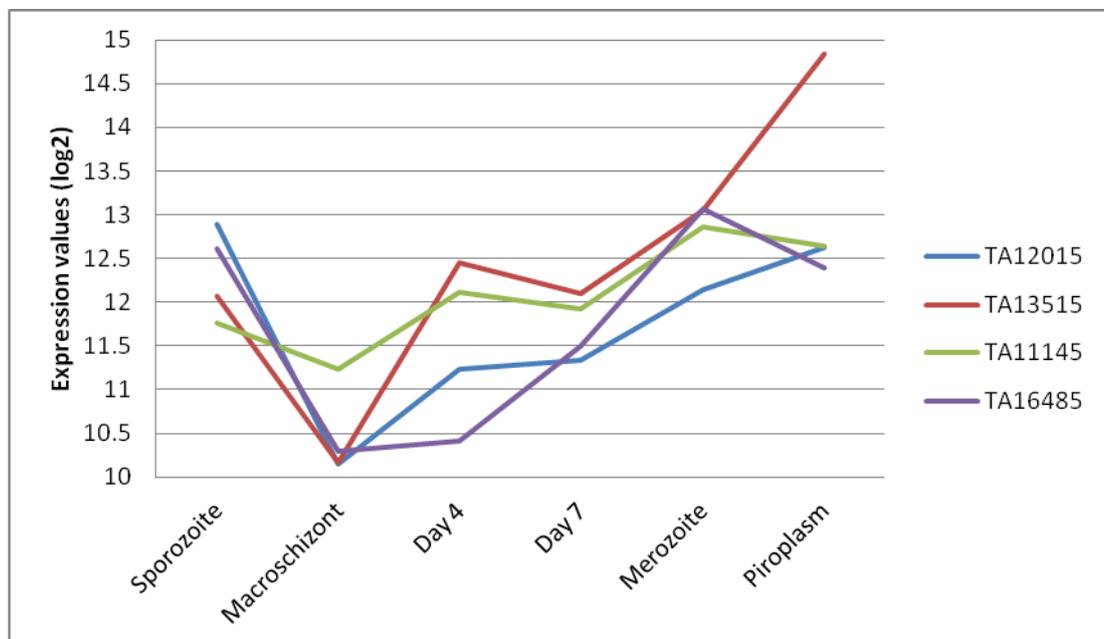


Figure 2.13. *T. annulata* temporal expression patterns of four most up-regulated potential transcription factors.

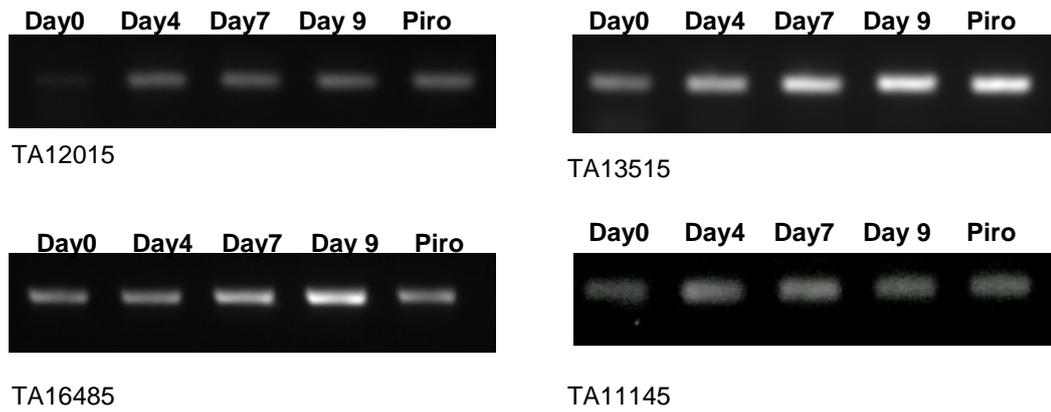


Figure 2.14. Validation of temporal profile of most up-regulated potential transcription factors by semi-quantitative RT-PCR.

Down-regulation of the AP2 domain factor TA13395 and the putative GATA TF (TA10735) was validated by semi-qRT-PCR (Figure 2.15, Figure 2.16). Only down regulation of the GATA factor was scored as significant by the microarray data but both genes showed evidence of reduced signal by RT-PCR with RNA from the Day 9 time point.

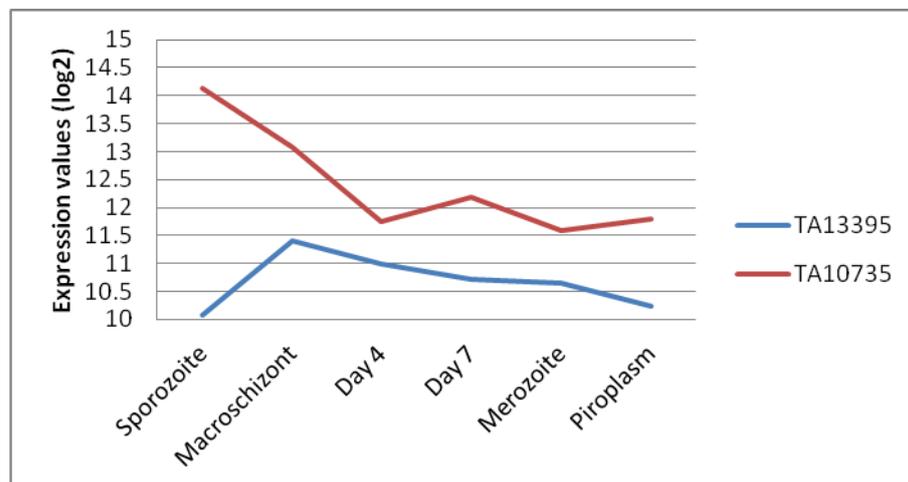


Figure 2.15. *T. annulata* temporal expression patterns of two most down-regulated potential transcription factors.

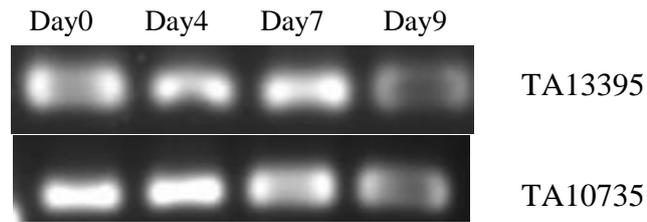


Figure 2.16. Validation of temporal profile of most down-regulated potential transcription factors by semi-quantitative RT-PCR.

2.3.8 Validation of expression pattern of selected up-regulated and down-regulated genes by qRT-PCR

To further validate the temporal expression profile (Figure 2.13) of two putative up-regulated ApiAP2 TFs, quantitative (q) RT-PCR was conducted. For validation by qRT-PCR, it was necessary to select control genes that are constitutively expressed. Heat shock 70 kDa protein (TA11610) gene and heat shock protein (Hsp) of 90kDa (TA10720) were selected as controls, as their expression profiles based on normalised microarray data did not reveal significant changes across the life cycle stages (Figure 2.17): this was confirmed by the semi quantitative PCR (see above).

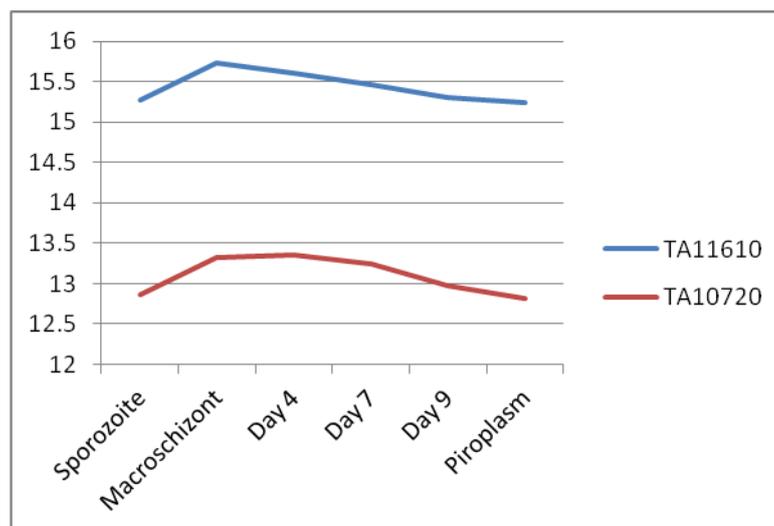


Fig.2.17. Temporal expression profile derived from microarray data of two heat shock proteins (TA10720 and TA11610) used as constitutive controls for qRT-PCR.

Primers were designed for TA13515 and TA11145 and qRT-PCR performed as described (section 2.2.5.3). As shown in Figure 2.18 qRT-PCR data obtained for normalised values of TA13515 showed evidence of elevated expression in relation to the calibrator time point, Day 0 (macroschizont). Relative to Day 0, a 3.64 log₂ fold change was obtained for Day 4 mRNA while Day 7 gave a 6.20 log₂ fold increase. Day 9 cultures differentiating to merozoites showed a 5.02 log₂ fold change, while a strong up-regulation in mRNA levels from merozoite to piroplasm stage was indicated (17.38 log₂ fold change, relative to the Day 0 time point). Statistical significance (p-value ≤ 0.05) was obtained for all stages except Day 4.

TA11145 showed a slightly different profile with a greater increase in mRNA levels from Day 4 through to cultures generating merozoites (Day 7 - 5.35 log₂ fold change and Day 9 - 4.5 log₂ fold change) and 5.99 log₂ fold change for the piroplasm. No statistically significant difference (p-value ≤ 0.05) between merozoite and piroplasm gene expression levels was found. Except for Day 4, statistical significance was obtained for all stages, relative to the macroschizont/Day 0 time point. It can be concluded that qRT-PCR validates that these two ApiAP2 genes are up-regulated during differentiation to the merozoite and the major up-regulation of the TA11145 gene occurs during merogony, whereas TA13515 ApiAP2 shows a greater elevation following differentiation to the piroplasm within the red cell.

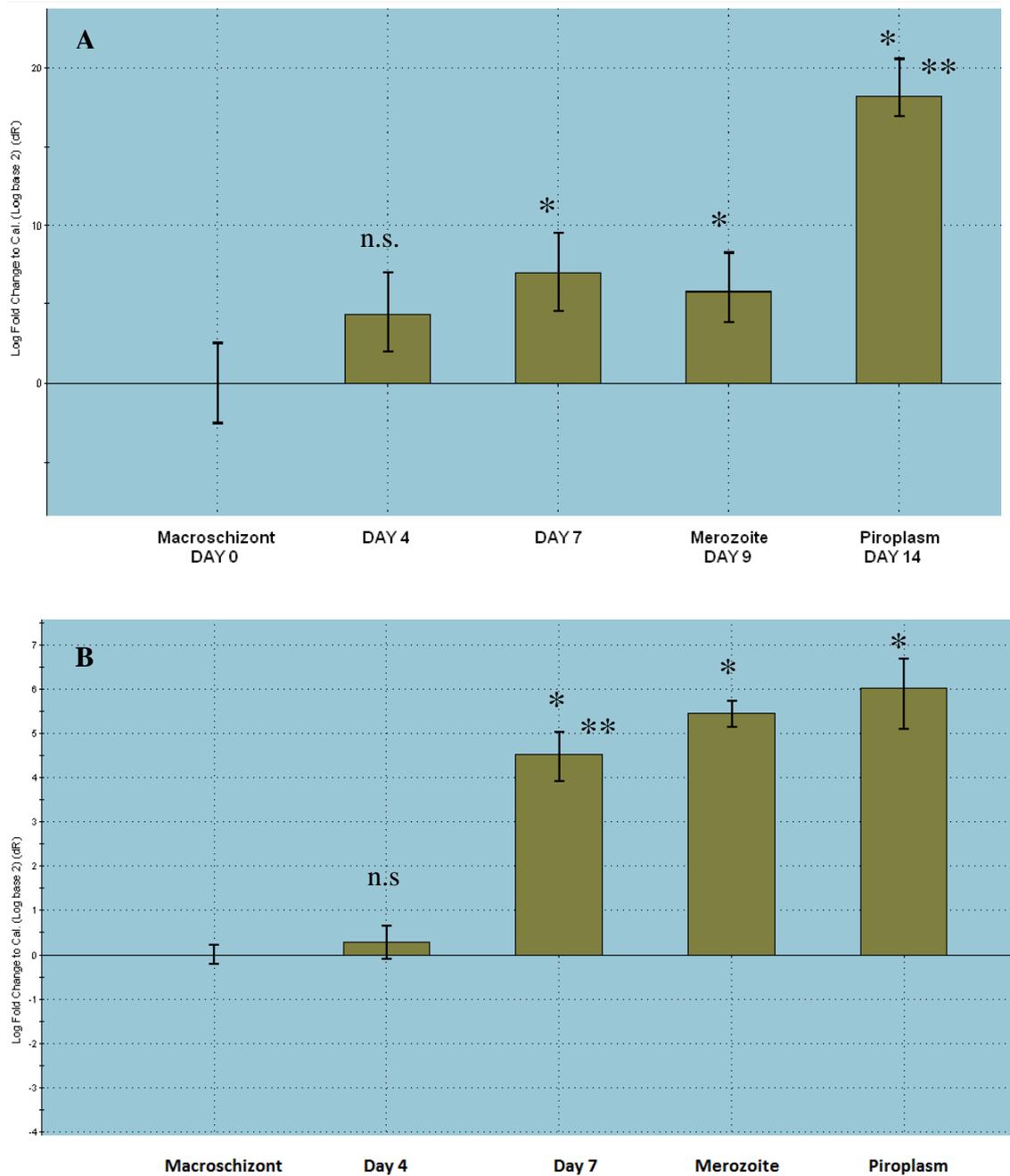


Figure 2.18. QRT-PCR results of A. TA13515 and B. TA11145: using RNA representing macroschizont (Day 0) cultures, and a differentiation time course to merozoite production (Day 7 and 9) and the piroplasm stage. Technical replicates values are normalised against Hsp70 (TA11610) and Hsp90 (TA10720) genes. The asterisk (*) above the bars indicates statistical significance (One Way ANOVA test and post hoc Dunnett test, $p\text{-value} \leq 0.05$) relative to macroschizont stage (*) and relative to its preceding stage (**). Whereas, (n.s.) indicates non-statistical difference of gene expression in particular stage relative to its preceding stage.

To validate genes that were indicated as down regulated during differentiation qRT-PCR was performed for TA15705 - Ta9 and TA10735 - (GATA TF, Figure 2.19). The results indicated a rapid, significant down-regulation between macroschizont and Day 4 (6.88 log₂ fold change reduction relative to macroschizont/Day 0, p-value ≤ 0.05) for TA10735, with maintenance of the reduced expression level through differentiation to merozoite (Day 7 - 7.14 log₂ fold reduction and Day 9 - 6.66 log₂ reduction, p-value ≤ 0.05). Gene TA15705 encoding the TA9 antigen (Figure 2.20.B) showed a slight up-regulation in the Day 4 differentiating culture (0.56 log₂ fold change relative to macroschizont/Day 0) followed by significant down-regulation at Day 7 (1.34 log₂ fold change reduction in expression level, p-value ≤ 0.05) and Day 9 differentiating cultures (3.16 log₂ fold change reduction, p-value ≤ 0.05).

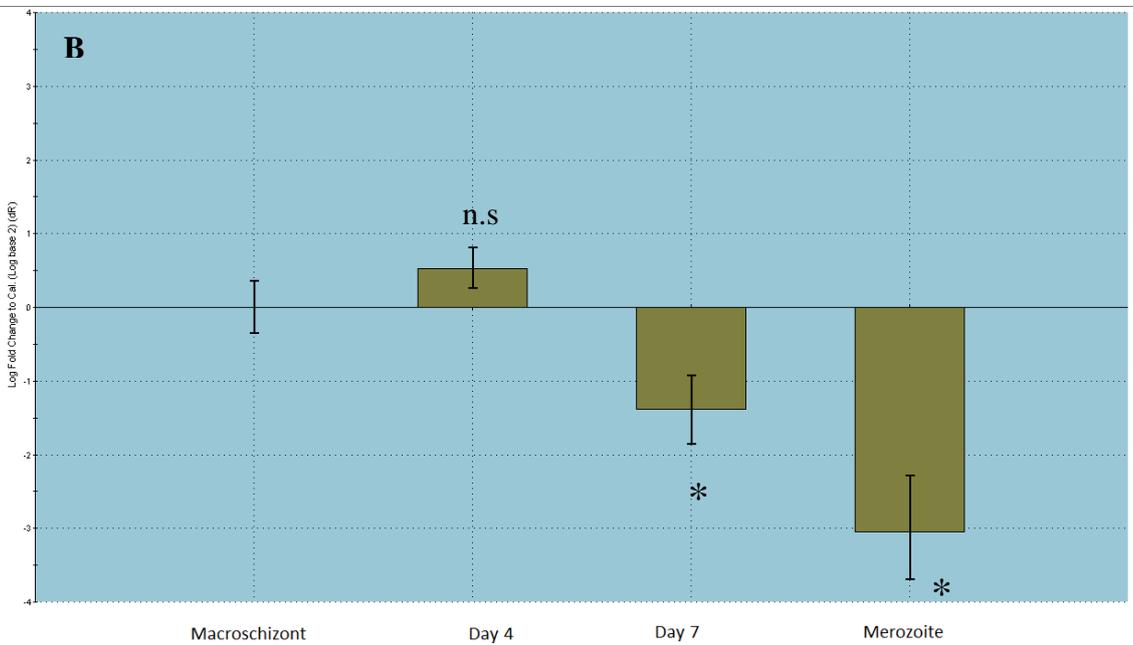
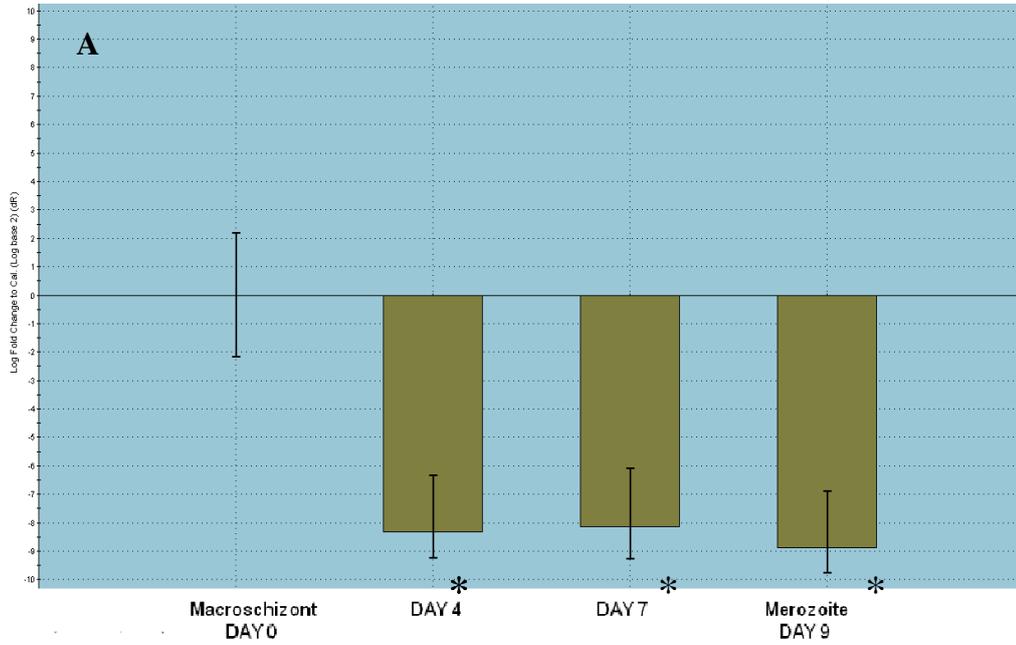


Figure 2.19. QRT-PCR results of A. TA10735 and B. TA15705 from macroschizont to merozoite stage. Technical replicates were normalised against Hsp70 (TA11610) gene. The asterisk (*) below the bars indicates the statistical significance of the result (One Way ANOVA test and post-hoc Dunnett test, $p \leq 0.05$) relative to macroschizont stage (*). Whereas (n.s.) indicates non-statistical difference of gene expression in particular stage relative to its preceding stage.

2.4 Discussion

The aim of the work presented in this chapter was to determine patterns of gene expression, at a whole genome level, during life cycle stage differentiation of *T. annulata*. This was achieved by construction of a tiled microarray (Weir, unpublished), hybridisation of RNA representing different life cycle stages, including an in vitro differentiation time course from macroschizont infected to merozoite producing cells. Initial analysis of the microarray data demonstrated that major differences in expression levels exist for a significant number of *Theileria* genes between days 4 and 7 of differentiation to the merozoite *in vitro*. Thus, up-regulation of genes encoding predicted and known merozoite polypeptides (e.g. rhoptry genes) and down-regulation of genes encoding polypeptides previously shown to be associated with the macroschizont stage (e.g. TashAT genes, Ta9) were identified.

The array data set indicated that approximately one fifth of the total number of *T. annulata* genes show significant changes in gene expression between macroschizont and merozoite life stages. This was illustrated using hierarchical clustering analysis and confirmed by Rank Product analysis of the normalised data set (W. Weir, unpublished). Interestingly, genes encoding polypeptides that are present at a high level in the merozoite stage are, in general, expressed at a low level in the macroschizont stage indicating that repression of gene expression in different stages is not absolute, based on designation of a baseline hybridization level to the microarray of 10 (\log_2 value).

The parasite has a relatively small genome (8.3 Mb) compared to the bovine and a large proportion of RNA extracted from infected leukocytes was of bovine origin. Despite contamination with host and tick vector RNA (sporozoite preps) >90% of parasite genes (including constitutively expressed genes) were indicated as expressed in each life cycle stage. Thus for the *T. annulata* candidate ApiAP2 transcription factors (see Table 2.1 and 2.2) all except TA10940 and TA16105 were predicted to be expressed at the RNA level during the macroschizont to merozoite/piroplasm phase of the life cycle. Furthermore, despite showing major up-regulation during differentiation to the merozoite/piroplasm all four up

regulated ApiAP2s showed a detectable level of expression in the macroschizont (see Figure 2.13-14). The results indicate that the conclusion of Shiels et al. (1992) that certain merozoite genes are expressed by the preceding macroschizont stage is applicable to the majority of merozoite genes, including those that may regulate this differentiation step. Similar results have been recorded for other Apicomplexa as changes in stage specific gene expression occur directly prior to a developmental commitment point. In *P. falciparum* sexual stage (gametocyte) gene expression was found to be present in the merozoite stage (Bruce et al., 1990) and similarly in *Eimeria bovis* merozoite genes were detected at the sporozoite life stage (Abrahamsen et al., 1993).

As shown by Table 2.2., RP analysis identified that a greater number of genes are up-regulated than down-regulated across the different stage differentiation steps analysed. Thus, 152 potential up-regulated genes and 115 down-regulated genes were captured (FDR<5%) between the macroschizont and merozoite stages, while using absolute fold change ≥ 2 denoted 361 up-regulated genes and 196 down-regulated genes. Such large-scale co-ordinated temporal regulation may involve a temporal switch in the level/activity of factors that control gene expression at the level of mRNA. Whether this is a single step or involves a cascade of regulation is not known. Differences in patterns of expression across stages and throughout the differentiation time course were observed for different genes at different data points. This suggests that a cascade involving multiple regulators is likely to occur. This postulation is supported by results from related studies in *Plasmodium* and *Toxoplasma* (Behnke et al., 2010; Painter et al., 2011), with the prediction of an interacting network of ApiAP2 DNA binding factors proposed for the intraerythrocytic cycle of *Plasmodium* (Campbell et al., 2010).

Rank Product analysis of particular genes showed that the most of the up-regulated genes from macroschizont to merozoite are hypothetical proteins, which prevented deeper analysis of a predicted function/role in the stage differentiation step. However, certain genes linked to known proteins were found and a selection those of interest are described. Rhoptry genes were clearly identified and their up-regulation was confirmed by semi-quantitative RT-PCR. This result was expected as rhoptry formation is known to occur during merogony and they are implicated in playing a central role in merozoite invasion

of the erythrocyte (Shaw and Tilney, 1992). Increased expression of cysteine proteinases genes was also indicated, confirming predictions from EST data (Weir, unpublished). This may not be surprising as they encode enzymes that degrade protein and could be required for morphological events or degradation of host cell proteins (Rosenthal et al., 1988). Surprisingly, the sporozoite surface antigen (SPAG) (Hall et al., 1992) was found to be among the top 100 up-regulated merozoite genes. This could indicate a gradual elevation through the life cycle in the tick to culminate at peak levels during sporogony. The gene encoding the macroschizont specific T cell antigen, Ta9 (TA15705) was found to be down-regulated from macroschizont to merozoite by array data and this was confirmed by semi-quantitative and qRT-PCR. Ta9 gene belongs to a small secretome gene family (5 genes), is recognized by T cells and is secreted into the host cytoplasm (MacHugh et al., 2011). The strong down regulation of this gene (also confirmed at the protein level by IFAT, Weir and Shiels unpublished) supports a role in transformation of the infected leukocyte. As predicted from previous work (Shiels et al., 1994; Shiels et al., 1998) the major merozoite surface antigen, Tams1 was found to be highly expressed in all life stages and showed up-regulation from macroschizont to merozoite/piroplasm stage. A Map2 kinase (TA21080) was also identified as up-regulated from macroschizont to merozoite stage. It belongs to the mitogen-activated protein kinase family involved in many cell processes, especially in directing cellular responses to different stress factors such as heat shock. These proteins are also implicated in proliferation, gene expression and differentiation and are crucial for maintenance of the cell (Schaeffer and Weber, 1999).

The largest *Theileria* gene families: subtelomeric SVSP genes, Tpr genes and TashAT genes were found to be expressed in macroschizont stage and then down-regulated in the next stages. This agrees with previous data in the literature (Swan et al., 2001; Shiels et al., 2004; Weir et al., 2010) and predictions from EST data (Weir, unpublished). Confirmation that TashATs and SVSPs are down-regulated during differentiation to the merozoite is of interest as the majority of these genes are postulated to contribute to the transformed phenotype, either by manipulation of host cell gene expression (TashATs) or the bovine immune response (SVSPs). Thus down-regulation may be required as host cell proliferation subsides and the leukocyte is destroyed (Shiels et al., 1992). In

addition, the tandem arrangement of related SVSP and TashAT genes could be useful to investigate any DNA motifs involved in coordinating their down-regulated expression.

Transcription in *T. annulata* has been shown to be a monocistronic process (Shiels et al., 2000) and, similar to *Plasmodium*, mechanisms that influence factor accessibility to promoter sites may play important role in regulation of gene expression. In *P. falciparum* it was shown that altered chromatin due to histone modification (de/acetylation and methylation) is linked with the transcriptional regulation of stage specific genes and progression of the *Plasmodium* life cycle (Luah et al., 2010). It was also proved that the addition of apicidin, a histone deacetylase inhibitor, can alter parasite differentiation status (Darkin-Rattray et al., 1996) and epigenetic control has been proposed to play role in apicomplexan differentiation steps. Epigenetic mechanisms may play a critical role especially in post-translational modifications of histone that affect the recruitment of chromatin-associated proteins or change interactions between histones and DNA (Coleman and Duraisingh, 2008). The presence in the *T. annulata* genome of histone deacetylases and methylases (Cock-Rada et al., 2012; Kinnaird et al., 2013) and also RNA-binding proteins (Pain et al., 2005) suggests that they are involved in regulation of gene expression. However, no statistically significant changes to expression levels of histone deacetylases and methylases between the macroschizont and merozoite stages were detected (data not shown) based on the D7 cell line microarray results. Thus, there is no evidence to indicate that a change in the level of histone modification enzymes operate to control gene expression in *T. annulata*, although a proteomic study comparing nuclear protein levels would be necessary to confirm this. However, based on the results from other studies it is likely that histone modification operates together with transcription factors. For example, treatment of *P. falciparum* blood-stage parasites with apidicin causes changes in expression of ApiAP2 genes, resulting in up-regulation of normally down-regulated genes (Chaal et al., 2010).

Most of the annotated *T. annulata* transcription factors (see Table 2.1 and 2.2) did not show major variation in expression levels, based on the array data, during macroschizont to merozoite differentiation. However, in several instances

some interesting alterations were observed. Thus, significant down-regulation of the TA10735 (a predicted GATA-type TF) was detected from the macroschizont to the merozoite stage. Strong down-regulation of TA10735 mRNA was also detected from sporozoite to day 4 of the differentiation time course and this low level of expression was then maintained through to the piroplasm stage. In total only three potential transcription factors showed statistically significant down-regulation (the other two containing an AP2 domain) suggesting that down-regulation of the GATA transcription factor (TA10735) is marked and significant. What role it plays in the life cycle is unclear but based on its profile it could be required for production of sporozoites or following sporozoite invasion of the leukocyte. The known function of GATA factors as regulators of developmental expression in higher eukaryotes (Briegleb et al., 1996; Rodriguez-Segui et al., 2012) points to a possible role in regulation of the parasite life cycle. In contrast, three potential TFs: TA06490 (SWI/SNF), TA06995 (ADA2 homologue) and TA17930 (hypothetical protein) showed a profile of elevated RNA expression in stages associated with intracellular infection of the bovine leukocyte. SWI/SNF complex is known to be involved in chromatin remodeling and to play important roles in gene expression in eucaryota (Sudarsanam and Winston, 2000; Tang et al., 2010) and ADA2 was found to regulate the histone acetyltransferase activity and promote transcriptional silencing (Hoke et al., 2008; Jacobson and Pillus, 2009). Clearly further work is required to investigate their function in determination of stage specific expression in *Theileria* parasites.

Based on the work of Balaji et al. (2008) that discovered the AP2 family of TFs in apicomplexan parasites, the identification of ApiAP2 transcription factors differentially regulated during differentiation steps of the *T. annulata* life cycle was of importance to this study. AP2 domain TFs are essential for differential expression in related *Plasmodium* (Yuda et al., 2009; Yuda et al., 2010) and most importantly, very recent work has shown that knock out of ApiAP2 genes in both *Toxoplasma* (Radke et al., 2013) and *Plasmodium* (Sinha et al., 2014; Kafssack et al., 2014) can alter the potential/efficiency of a stage differentiation event and are involved in sexual determination in *Plasmodium*.

Analysis of the temporal expression profile of the 22 *T. annulata* ApiAP2 genes demonstrated that the majority of them are expressed at the RNA level above

the background threshold during macroschizont to merozoite stage transition step. However, one ApiAP2 gene (TA10940) might be considered as sporozoite or tick stage specific as its expression values during macroschizont to merozoite/piroplasm stage are below the cut-off threshold (expression value=10). As microarray assay is not always 100% reliable with respect to mRNA expression of particular genes due to a limited number of replicates and diversity of experimental conditions, it would be of interest to confirm a possible sporozoite or tick stage specificity of this ApiAP2 gene.

Most of the ApiAP2s do not show significant changes in expression level but this does not rule out a role in control of gene expression or in stage differentiation. Modest changes in expression values may be missed by the application of the significance cut-off applied to microarray data - and to some extent this was indicated, as microarray data relative to qRT-PCR gave smaller changes in stage expression levels. In addition, in other studies relatively small changes in expression levels of ApiAP2 during stage differentiation events were recorded using array data (Young et al., 2005; Silvestrini et al., 2005; Campbell et al., 2010).

Four *T. annulata* ApiAP2 genes (TA13515, TA16485, TA11145 and TA12015) showed up-regulated expression relative to the rest of the family (see Figure 2.9.A). This suggests they may be functionally required by the merozoite or piroplasm stages and could play a role in stage differentiation if they act to specifically regulate a subset of target genes. Three of these genes (TA13515, TA12015, TA16485) were in the 100 most up-regulated genes from macroschizont to merozoite and showed a marked reduction in RNA levels in the macroschizont stage relative to all other stages. While these profiles of ApiAP2 are of most interest, it remains to be established whether they are linked to or required for regulation of *T. annulata* stage differentiation. In *Plasmodium*, the expression profiles of putative co-expressed targets are both positively and negatively correlated to the expression profile of the corresponding ApiAP2 gene, indicating that they function either as activators or repressors of transcription (Campbell et al., 2010; Painter et al., 2011). The differentiation expression pattern of the four *Theileria* up-regulated ApiAP2 genes suggests that they could mediate transcriptional regulation of stage specific genes. Thus strong correlation of

potential target genes with a particular ApiAP2 gene might suggest regulation by this factor. Thus, for example the TA13515 AP2 TF could control target genes that are regulated during merogony or more likely following differentiation to the piroplasm, since this transition shows the highest level of increase in TA13515 expression. Semi-quantitative and quantitative RT-PCR confirmed the up-regulated expression pattern of TA13515 and also the TA11145 gene. The TA11145 profile was of particular interest as it was proposed that regulators of merozoite gene expression would be expressed in the macroschizont stage and show an increase prior to a commitment point leading to auto-regulatory control (Shiels et al., 2000). TA11145 expression is predicted to be higher than TA13515 in the macroschizont stage based on microarray data and qRT-PCR results. Expression of TA11145 by qRT-PCR data showed a small increase between macroschizont stage and Day 4, followed by a more significant increase between Day 4 and the Day 7 and Day 9 points where merozoite production has occurred. Following merozoite production however, TA11145 mRNA did not show a significant elevation to the piroplasm stage, whereas TA13515 expression probably peaks in the piroplasm or early phases of the parasite life cycle within the tick (gametocyte to kinete - not determined). Work of Sinha et al. (2014) and Kafsack et al. (2014) revealed that one of the members of *P. berghei* ApiAP2 family, PbAP2-G (an orthologue of *P. falciparum* PfAP2-G - PFL1085w/PF3D7_1222600) is essential for the commitment of asexually replicating forms to sexual development. Moreover, this work also confirmed the postulation that a cascade of ApiAP2 proteins are involved in commitment to the production and maturation of gametocytes. Based on the phylogeny generated in the study of Balaji et al. (2005) and confirmed by analysis conducted in this study the *T. annulata* AP2 domain-encoding gene TA13515 is an ortholog of PbAP2-G. Thus, the expression profile indicating significant up-regulation of TA13515 in the late piroplasm stage points to the possibility of an orthologous role in determination of commitment to gametocytogenesis/ sexual phase of the *T. annulata* life cycle.

The presence of the four up-regulated ApiAP2 genes during the macroschizont to merozoite stage transition indicates that the genes targeted by these predicted DNA binding factors differ. This may point to temporal control of gene expression during differentiation or a different role such as enhancer or

repressor. To study this further the binding sites of the different *T. annulata* AP2 domains requires identification. Conservation of binding sites across *Apicomplexa* species (Guo and Silva, 2008; Essien and Stoeckert, 2010), in particular ApiAP2 binding specificities across the *Plasmodium* genus (Painter et al., 2010) has been described. Thus, utilization of orthologous ApiAP2 DNA binding motifs described for *Plasmodium* would be a good starting point for analysis of *Theileria* AP2 binding sites and target genes possessing these sites. A general search for shared conserved DNA motifs associated with the differentially regulated genes should be performed to identify any putative motifs not recognized by selected AP2 ortholog domains. Identifying genes encoding potential DNA binding factors together with sets of genes, that possess, in their promoter region, the motif the factor is predicted to bind to suggests that the factor may regulate these genes. Correlation (both negative and positive) of expression patterns of the binding factor and potential target genes increase this probability. It is also relevant to see if the motif the factor is predicted to bind to is present in its own upstream region, implying auto-regulation a key element of the stochastic model of stage differentiation.

In summary, the work in this chapter has provided a data set of genes that show significant changes to expression levels as the parasite undergoes transition between life stages. It has also highlighted changes to expression of important potential *Theileria* transcription factors/DNA binding proteins. This provided data sets and candidate factors for further study and set a key goal of: identification of nucleotide motifs in the intergenic regions of differentially expressed genes that could be bound by the DNA binding proteins; and investigation of potential target motifs bound by differentially regulated putative ApiAP2 transcription factors in *T. annulata*.

Chapter 3

Identification and analysis of *T. annulata* cis-acting promoter elements

3 Identification and analysis of *T. annulata* cis-acting promoter elements

3.1 Introduction

Apicomplexan species have reduced complexity in their transcriptional apparatus in comparison to other eukaryotic organisms, hence different mechanisms of transcriptional regulation are predicted for these parasites, including *Theileria* species. It has been demonstrated, in many eukaryotic systems, that changes to the profile of gene expression following cellular differentiation is mediated by interactions between polypeptide factors and short nucleotide motifs located upstream of the transcription initiation point (Latchman, 1991; Blau, 1992; Thomas and Chiang, 2006; Lengar and Joshi, 2009). This type of mechanism is supported by the highly compact genome of *T. annulata*, with non-coding regions forming around 30% of the genome and short intergenic regions (IGRs) of, on average, 400 base pairs (Guo and Silva, 2008). Intergenic regions are likely to contain cis-regulatory elements (CREs), non-coding DNA motifs of six to eight nucleotides in length which act as transcription factors binding motifs (TFBMs). The sequence pattern of the DNA motif is very important, as it determines the binding affinity of transcription factors (TFs) (Zhang et al., 2003). The effect of binding can be either positive (enhancement of transcription) or negative (repression of transcription).

Conservation of regulatory regions between species implies an important biological role (Lee et al., 2005). Interspecies comparison of non-coding regions of yeast genomes revealed that conserved DNA motif sequences of orthologous genes act as binding sites (Kellis et al., 2003; Doniger et al., 2005). A considerable degree of variability across regulatory regions between species may arise over time due to evolutionary change that can affect non-coding regions with a higher frequency than coding sequences (Dermitzakis and Clark, 2002). Mutations in cis-element sequences may not block binding completely, as there is an amount of flexibility at the motif level and not all positions in the binding site equally contribute to binding and therefore a transcription factor (TF) may still bind to a mutated cis-element but with lower affinity (Benos et al., 2002). In addition to the important role of hydrogen bonding, the efficacy of sequence-specific binding is also influenced by interactions with other proteins and the

chromatin state/structure (Benos et al., 2002). Presently, the most common cause of phenotypic divergence in transcriptional control is thought to be mutations affecting DNA motif sequences of cis-elements rather than trans-elements encoding the transcription factors proteins binding to these sequences (Carroll, 2008; Stern and Orgogozo, 2008). Mutations arising in non-coding regions have been shown to influence gene expression by changing the way transcription factors bind to their target sequence, leading to up-regulation, down-regulation or complete deactivation of the gene (Wittkopp and Kalay, 2012).

The most well characterised types of cis-regulatory elements are enhancers and promoters. Promoters are short (~40bp) regions of DNA which are located upstream of the transcription start site (Ayoubi, 1996; Butler, 2002). By binding to this region, TFs induce transcription of the downstream gene. Apicomplexan species, however, lack typical eukaryotic promoter elements such as the TATA-box and CAAT-box (Millitello et al., 2005) as well as the corresponding transcriptional factors such as TATA-binding proteins and TFIIA (Meissner and Soldati, 2005). In contrast, enhancers are regions of DNA that act by influencing the transcription of the gene. They can be found upstream and downstream of the associated gene, within introns or can be located far away from the gene they regulate. Multiple enhancers can act together to regulate transcription of one gene (Wittkopp and Kalay, 2012). Enhancers are thought to be responsible for cis-regulatory divergence as their sequences are more variable between species (Wray, 2007) than promoters. Some enhancers have overlapping functions hence mutation in single instance may only have a limited effect on gene expression (Hong et al., 2008).

The availability of an increasing number of sequenced genomes together with publication of various experimental studies have facilitated the development of bioinformatic techniques to investigate DNA sequence motifs responsible for transcriptional regulation. Using web-based bioinformatics tools, searches at genomic-level sequence datasets can be performed for DNA motifs that act as known factor recognition sites. Alternatively, genomic sequences can be screened for novel motifs that have been conserved, and repeatedly used, through evolution by scanning regions upstream of potentially co-regulated

genes. Strong conservation of the motif provides additional evidence of its functional relevance. In Apicomplexa the process of transcription is considered to operate in a monocistronic manner (Iengar and Yoshi, 2009). Evidence that this is the case for *T. annulata* has been generated (Shiels et al., 2000) but only limited information on DNA regulatory motifs and TFs of stage-specific *T. annulata* genes has been obtained to date. In the study of Shiels et al. (2000) a 9bp core binding site 5'-TTTGTAGGG-3' was identified in the upstream region of the merozoite surface antigen gene, Tams1. Nucleotides flanking this region included a CACACA motif that is understood to stabilise the formation of DNA motif-factor complexes.

Guo and Silva (2008) demonstrated three putative motifs present in hundreds of copies throughout the genomes of *T. parva* and *T. annulata* (Figure 3.1). The motifs 'TCCCAT' and 'GATTCCA' in *T. parva* were found to be located 60 nucleotides upstream of the transcriptional start site, which suggests that they may play a role as transcription factor binding sites in each species. In contrast, the motif 'TGTGT' showed no fixed localisation relative to the transcriptional start site and was found throughout non-coding regions of the genome. This motif was also described for *Plasmodium* species (Young et al., 2008) and, similar to *T. parva*, was shown to have a widespread distribution in the genome.

<i>T. parva</i> Motifs	Similar Motifs in <i>T. annulata</i>	Best Matching Known Motifs	Functional Enrichment
		Name (Source) Logo E-value	
		MZF1_1-4 (JASPAR) 	3.4e-04 Telomeric ORF Signal Peptide Protein Synthesis
		NAPINMOTIF (PLACE) 	7.1e-10 /
		NF-kappaB M00051 (TRANSFAC) 	4.2e-05 Protein Fate

Figure 3.1. Top three statistically significant MEME-derived motifs in *T. parva* and the most similar motifs in *T. annulata* (Guo and Silva, 2008) (two left columns). Functional annotations enriched in downstream genes of each motif in *T. parva* are shown on the right. The best matches to conserved *T. parva* motifs are shown in the centre (motif name and database source, sequence logo and E-value, which represents a relative measure of similarity between two motifs based on simulated position specific score matrix models).

Motif comparison analysis also showed that the top three statistically significant motifs in *T. parva* are almost identical to motif 1, 2 and 4 from the top of the list of motifs identified for *T. annulata* (Guo and Silva, 2008). The 'TCCCAT' motif was found to be enriched in *T. parva* promoter regions of genes encoding

signal peptide-containing proteins and also of telomeric open reading frames (ORFs) encoding polymorphic secretory gene families (Bishop et al., 2000). This motif is similar to the G-box motif previously identified for *Plasmodium* species (Millitello et al., 2004; Sunil et al., 2008, Young et al., 2008) in upstream regions of ribosomal and heat shock genes (Millitello et al., 2004). It has been shown in *P. falciparum* that some ApiAP2 domains differentially expressed in specific developmental stages have high specificity for unique DNA sequence motifs found in the upstream regions of some genes (Campbell et al., 2010) (Figure 3.2) and this includes a G-box upstream of the Hsp genes. A G-box motif was also identified as a binding site for a *C. parvum* ApiAP2 transcription factor, Cgd8_810 (Oberstaller et al., 2013).

In *P. falciparum* the 'CACA' motif was observed in the upstream regions of genes from the DNA replication and chromosome cycle cluster (Young et al., 2008) and variations of this motif have been found to act as binding sites for *P. falciparum* AP2 domains (Campbell et al., 2010) (Figure 3.2). This motif was also found in the upstream regions of *Toxoplasma* genes (Gross et al., 1996; Bohne et al., 1997). The third motif identified by Guo and Silva (2008), 'GATTCCA', was found to be similar to the binding site of the NF- κ B transcription factor known to be involved in host cell transformation by *Theileria* species (Heussler et al., 2002). Localisation of these three motifs in 5' untranslated regions (UTRs) more frequently than in the other regions is consistent with the presence of regulatory motifs in the upstream region of target genes. Moreover, two putative *T. parva* motifs were located just upstream of transcriptional start sites and were preferentially associated with specific protein functions, supporting the hypothesis of participation in transcriptional regulation of this parasite (Guo and Silva, 2008). Whether this finding is associated with the function of NF- κ B in host cell transformation remains unknown, but appears unlikely.

Finding an enrichment of the same motifs in closely related *Theileria* species suggests gene expression regulation by similar orthologous transcription factors. Examination of orthologous clusters of putative families of transcription factors that contain ApiAP2 domains indicated a close relationship between *Plasmodium* and *Theileria* ApiAP2 genes within the Apicomplexa (Figure 3.3) and suggests they may bind similar DNA motifs. They share 16 predicted domains from at least

15 distinct proteins; while *Cryptosporidium* and *Plasmodium* share only 11 ApiAP2 domains from nine proteins. It is likely that the common ancestor of apicomplexans possessed at least nine members of an ApiAP2 family, which appears to have proliferated further through independent gene duplication as the different apicomplexan lineages emerged (Balaji et al., 2005). *Toxoplasma* showed a distinctly higher number of ApiAP2 genes than other apicomplexans and also larger genome, which may indicate a higher degree of specific transcriptional regulation (Iyer et al., 2008).

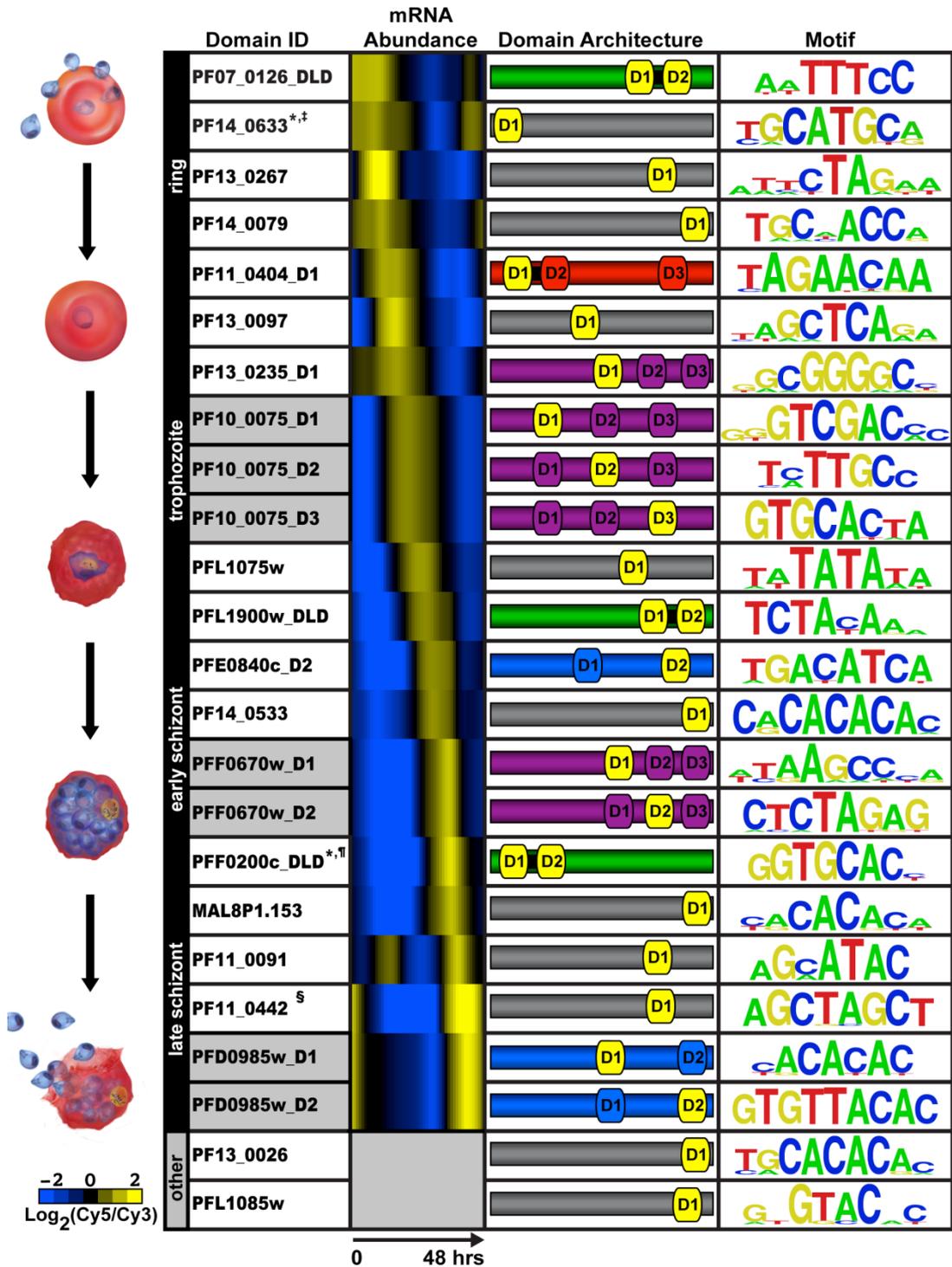


Figure 3.2. *P. falciparum* ApiAP2 domains and their corresponding target DNA motifs. In addition: relative mRNA expression profiles during the intraerythrocytic developmental cycle (IDC) (genes not expressed during the IDC are in grey), AP2 domain(s) architecture and location in the protein (Campbell et al., 2010).

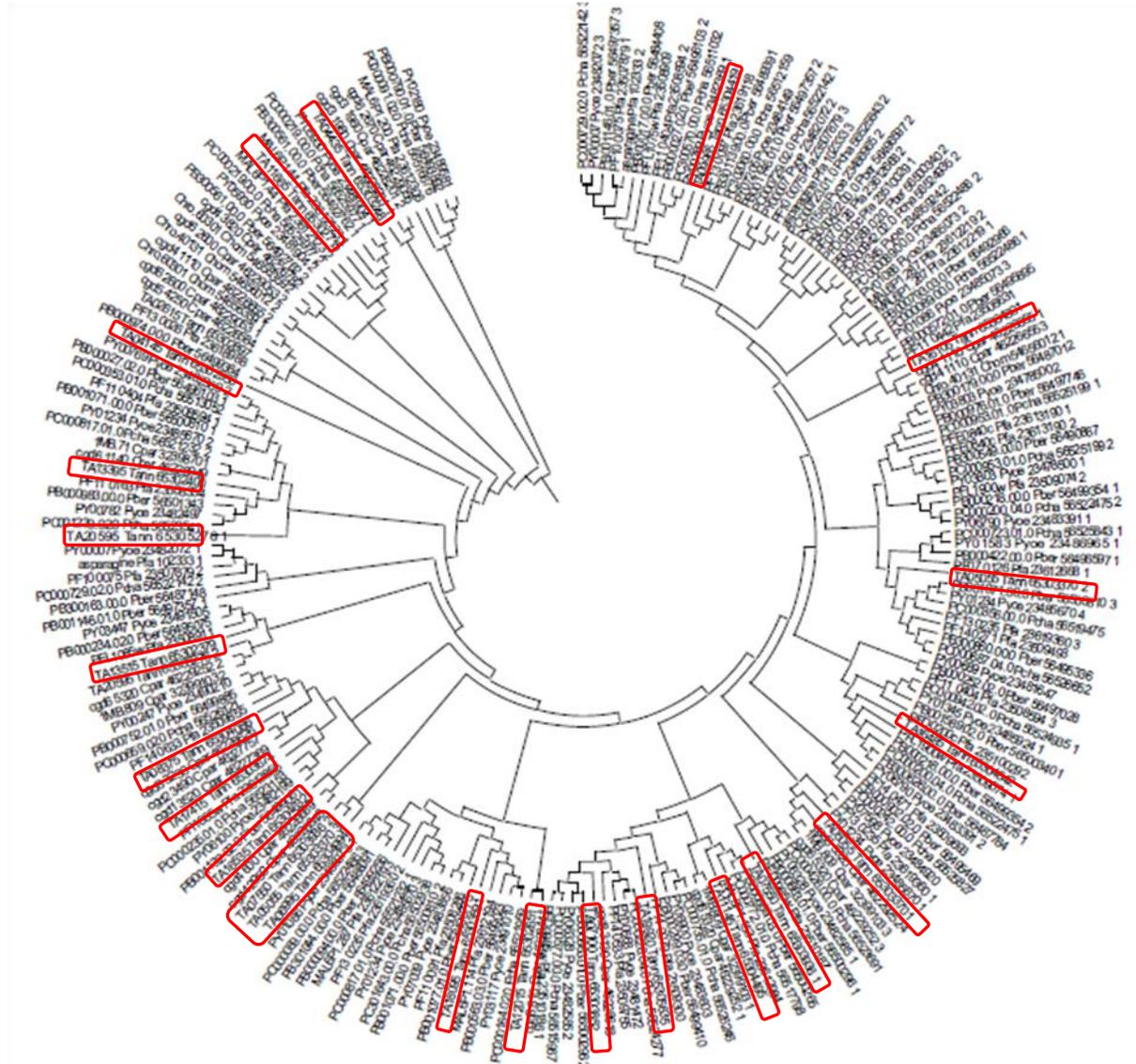


Figure 3.3. UPGMA tree of the ApiAP2 domain in Apicomplexa (*Theileria* ApiAP2 domain - highlighted in red) (Balaji et al., 2005)

Proteins functioning as regulatory factors involved in the control of *Theileria* gene expression during differentiation to the merozoite still await identification. It is likely that genes with similar mRNA expression profiles (co-expressed genes) are regulated via the same mechanism (co-regulated). Such co-expressed genes are likely to have conserved DNA binding motifs in their promoter region. It was shown that some ApiAP2 domains differentially expressed in specific developmental stages have a high specificity to unique DNA sequence motifs found in the upstream region of the transcription initiation site of some genes (Campbell et al., 2010) (Figure 3.2). Individual ApiAP2 genes showed slight temporal differences in their expression patterns within stage-specific guilds and comparison of these expression patterns with the rest of the gene set expressed by individual stages might indicate target genes that they regulate. Thus genes showing strongly correlated positive or negative expression with a particular guild of ApiAP2 genes might be under the control of the products of that guild (Balaji et al., 2005). A number of studies have indicated that ApiAP2-related transcription factors are widely used for gene regulation throughout the *Plasmodium* (Yuda et al., 2010; Flueck et al., 2010; Sinha et al., 2014) and *Toxoplasma* (Radke et al., 2013) life-cycle. For example, AP2-Sp (Pf14_0633; Pbanka_132980) is a major TF expressed from the late oocyst stage to the sporozoite stage and regulates gene expression in the sporozoite by binding to the 'GCATGCA' DNA motif (De Silva et al., 2008; Lindner et al., 2010; Painter et al., 2011). Disruption of this gene does result in a loss of sporozoite formation but does not affect parasite replication in the erythrocyte (Yuda et al., 2010).

Computational methods used by the studies described above are an important tool for further investigation of potential DNA sequence motifs involved in the regulation of gene expression in *T. annulata*. Moreover, identification of these motifs and the factors that bind to them may provide a deeper insight into mechanisms that regulate stage differentiation of this parasite. The work described in this chapter concentrates on identifying DNA motifs in the genome of *T. annulata* and investigating their potential role in regulation of specific subsets of genes differentially expressed during stage differentiation. In addition, it was important to gain insight into motifs recognised by the four *T. annulata* genes encoding putative ApiAP2 transcription factors (TA13515, TA11145, TA12015, and TA16485) up-regulated during merogony (see Chapter 2).

Conservation of these *T. annulata* ApiAP2 factors across other *Theileria* species (*T. parva* and *T. orientalis*) and other Apicomplexan genera may point to the possibility that the motifs they recognise are conserved. Thus, an aim of the work in this chapter was to determine if AP2 domains represented by these up-regulated ApiAP2 genes in *T. annulata* show strong conservation across related species and genera. If so, study of the occurrence of binding ApiAP2 sites previously identified in *P. falciparum* and *P. berghei* (Campbell et al., 2010) (GTGTAC, CACAC, TCTACA) and *C. parvum* orthologue (Oberstaller, 2013) (G-box) would then be conducted to determine if there is an association of predicted motif with stage-specific regulation of gene expression in *T. annulata*. It was hoped that would provide insight into the regulation of ApiAP2 factors and target genes during stage differentiation and lead to further understanding of the mechanisms operating in a stochastic model of stage differentiation.

3.2 Methods

3.2.1 Motif discovery in upstream regions of *T. annulata* co-expressed genes and gene families

The small size of regulatory sequences and often a lack of positional conservation makes it difficult to identify cis-acting elements using local alignment approaches such as BLAST. With the increased availability of whole genome sequences, several algorithms have been developed to detect over-represented DNA motifs in the non-coding sequences either *de novo* or searching for known motifs (Tompa et al., 2005). One of the most popular and widely used tools for discovering and analysing motifs in sets of DNA or protein sequences is MEME (Multiple Expectation Maximization for Elicitation of Motifs), version 4.6.1 (Bailey and Elkan, 1994). MEME uses a deterministic approach to build models of sequence motifs assigning sites to the motif and building on this assignment as more occurrences of the same cis-element are found (D'Haeseleer, 2006). It operates on the basis of an expectation maximization (EM) algorithm and identifies shared motifs and their positions in unaligned sequences. There are three options for the motif search: exactly one motif per sequence, one or zero motifs per sequence or any number of motifs per sequence. Individual motifs provided by MEME do not contain gaps.

Input sequences predicted to share (unknown) motifs were prepared by extracting intergenic sequences upstream and downstream of the coding sequence (CDS) of genes of interest using the GeneDB database (Hertz-Fowler et al., 2004). Since transcription start sites have not been determined for all the *T. annulata* genes, intergenic regions based on the position of the translation start site (TSS) - ATG start codon, to the next coding sequence were extracted. Sequences of upstream regions from co-expressed genes (up- and down-regulated from macroschizont to merozoite) and gene families in FASTA format were submitted to the MEME (www.meme.nbcr.net) to look for repeated un-gapped sequence patterns which may represent potential binding sites for the transcription factors. One big advantage of MEME software motif prediction is the possibility to adapt motif length (Hu et al., 2005) in searches. Motif searches were performed using a motif length of between 5 and 8bp, 8 and 12bp and 8

and 20bp and the zoops (zero or one occurrence per sequence) option to increase probability of finding significant motifs.

The MEME results are presented as a summary table showing motif sequence logo, length, number of sites contributing to the construction of the motif and statistical significance. The statistical significance of the motif is presented as an E-value based on an estimation of the expected number of motifs with the given log likelihood ratio and with the same width and site count that could be expected in a similarly sized set of random sequences. Generated sequence logos represent the probability of each possible letter occurring at each position of the motif. The height of individual letters in the logo represents the probability of the letter at that position, multiplied by the total information content of the motif. Nucleotides were colored as follows: Adenine - red, Guanine - orange, Cytosine - blue, and Thymine - green. Additionally, occurrences (sites) of the motif across the dataset were presented as an alignment and sequences preceding and following each motif are shown. Motif sites were listed in order of increasing statistical significance (p-value) which gives the probability of a random sequence motif generated from the background letter frequencies having the same match score or higher. Each site was identified by the name of the gene where the upstream region occurs, the strand (sense and anti-sense in reference to the orientation of the CDS) and position in the sequence where the motif starts. For the negative strand the start position of the motif is actually the position on the positive strand where the motif ends. Statistically significant motifs were presented as coloured blocks on a line diagram displaying all the sites in the IGR where the DNA motif was identified on the positive or negative strand. The p-value of each occurrence and a sequence position scale is additionally shown in each diagram.

3.2.2 Cross-species MEME analysis of the TashAT cluster

MEME analysis was additionally performed with *T. annulata* and *T. parva* TashAT family orthologues, as looking across species in families of genes that have diverged may support any functional significance ascribed to any identified motifs. Cross-species MEME analysis of the upstream regions of 17 *T. annulata* TashAT genes and 20 *T. parva* TpHN orthologous genes was previously performed (Pain and Shiels, unpublished) as above. Motif searching on both strands was

undertaken using a motif length of between 6 and 50bp and zoops (zero or one occurrence per sequence). The obtained motifs and their localisation in the analysed sequences are presented in the results section.

3.2.2.1 Analysis of the intergenic region of TA03110 (Tash-a) across *T. annulata*, *T. parva* and *T. orientalis*

TA03110 is the only gene in the TashAT cluster which is conserved in the non-transforming *Theileria* species, *T. orientalis*, and is up-regulated during merogony, based on microarray data and IFAT (Hayashida et al., 2013). To determine if the 4C-box motif (see following section) is associated with down-regulation of genes in the TashAT cluster, analysis of the 5' intergenic region of the gene encoding Tash-a (TA03110) and its orthologues in *T. parva* and *T. orientalis* was performed. Intergenic regions upstream of the ATG start codon to the margin of the neighbouring coding sequence of *T. annulata* and *T. parva* were extracted from the GeneDB database. The upstream sequence of *T. orientalis* was extracted manually using Artemis: Genome Browser and Annotation Tool (Rutherford et al., 2000). A motif search was carried out using MEME with a motif length of between 6 and 14bp and zoops (zero or one occurrence per sequence).

3.2.3 Statistical enrichment analysis of 4C-box and 5C-box motifs within *T. annulata* intergenic regions

Analysis of DNA motifs derived by MEME software revealed 4C-box (TCCCCAT) and 5C-box (TCCCCCAT) elements present in the upstream regions of TashAT and SVSP family genes, respectively. Based on the hypothesis that genes from the same family which are co-expressed should also be co-regulated and may possess common upstream regulatory elements, an additional motif enrichment analysis was performed for these two putative cis-elements to confirm their statistical significance. Based on an average size of *T. annulata* intergenic regions, a length restriction of 400 bp was used in this analysis. Additionally motifs were searched for 100bp upstream of the ATG start codon using the DNA Motif Pattern Search option in PiroplasmaDB[®] (www.piroplasmadb.org). A motif enrichment value was calculated using a custom Perl script written by Dr William Weir (University of

Glasgow), which counts the occurrence of each motif in upstream regions of gene clusters or gene families and compares it with the number of occurrences in the remainder of the genome. A P-value was calculated using a two-tailed Fisher's exact test (<http://www.graphpad.com/quickcalcs/contingency1>) and statistically significant results (p-value ≤ 0.05) highlighted as significant associations between categorical variables.

3.2.4 ApiAP2 domain conservation analysis across *Theileria* and *Plasmodium* species

Based on expression data from Chapter 2, four up-regulated macroschizont to merozoite *T. annulata* genes (TA13515, TA11145, TA16485 and TA12015) were selected for further detailed analysis. Orthologues in *T. parva* and *T. orientalis* were identified by Protein BLAST search (www.ncbi.nlm.nih.gov/BLAST) and ApiAP2 domain regions extracted. *Theileria annulata* ApiAP2s domain boundaries were defined as in Balaji et al. (2008) and confirmed using the Pfam database (<http://pfam.sanger.ac.uk>). In addition, Protein Blast searches for these four ApiAP2s was performed and a list of putative Apicomplexan orthologues with a minimum of 50% sequence identity with *T. annulata* ApiAP2 domain sequence was generated.

Multiple sequence alignment across the three *Theileria* species was performed for the four up-regulated *T. annulata* ApiAP2 using ClustalW online software (www.ebi.ac.uk/Tools/msa/clustalw2/) and edited with Jalview (<http://www.jalview.org>). Alignments were also performed with the *P. falciparum* ApiAP2s domain orthologues for TA11145, TA13515 and TA16485 to visualise domain conservation. As no close *P. falciparum* orthologue was found for TA12015, it was decided to use the *C. parvum* orthologue instead, as its DNA target motif has been established (Oberstaller et al., 2013). An alignment was performed for all TaApiAP2s using T-coffee software (www.ebi.ac.uk/Tools/msa/tcoffee/) and illustrates general conservation of the ApiAP2 domain in *T. annulata*.

3.2.5 Distribution and enrichment analysis of *Plasmodium* motifs identified in *Theileria* genome

Plasmodium falciparum ApiAP2 domain DNA target motifs were selected for further analysis to screen for potential conservation of ApiAP2 DNA binding sites in promoter regions of *Theileria* genes. *Theileria annulata* genes with the *Plasmodium* DNA binding motif in their upstream region were selected using PiroplasmaDB DNA motif pattern search (<http://piroplasmadb.org/piro>). Due to the relatively small *T. annulata* genome and short (~400bp) intergenic regions (Guo and Silva, 2008), a length restriction of 400bp regions upstream of the ATG translation start site was chosen to perform motif searches. To investigate motif distribution in the *Theileria* genome, motifs were searched within 0-100bp, 101-200bp, 201-300bp and 301-400bp regions upstream the translation start site and also within the first 200bp of the protein coding region. To identify potential target genes of selected TaApiAP2 domains (TA13515, TA11145, TA16485, TA12015), groups of genes co-expressed with the ApiAP2 factor across multiple differentiation time-points and known stage-regulated gene families were screened for the predicted motifs. The obtained data was exported as an Excel file and motif distribution data tabulated. A motif enrichment value was calculated using the custom Perl script as described above and P-values calculated in a similar manner. Analysis was performed for each *Plasmodium* motif (Campbell et al., 2010) separately and motif over-representation determined.

3.2.6 Motif enrichment analysis of the upstream regions of ApiAP2 genes in *Theileria*

In order to investigate the issue of TaApiAP2 auto-regulation, a sequence alignment of intergenic regions of the four up-regulated ApiAP2 *T. annulata* and their *T. parva* orthologues was performed using ClustalW software (www.ebi.ac.uk/Tools/msa/clustalw2). The alignment was then analysed with Jalview software (<http://www.jalview.org>) to check for motifs potentially missed by MEME. Upstream regions of selected TaApiAP2s (TA13515, TA11145, TA12015, TA16485) were searched for respective DNA binding motifs (GTGTAC, CACACA/ACACAC, G-box/C-box, TCTACA). These had previously been identified

for orthologous genes in *P. falciparum* (Figure 3.2) and *C. parvum* (Oberstaller et al., 2013).

3.2.7 Temporal expression pattern analysis of ApiAP2 genes and their potential target genes

Correlation (either positive or negative) of expression pattern of genes which possess the same binding motif in their upstream region with the expression pattern displayed by particular ApiAP2 genes (predicted to bind the motif) could indicate regulation by these putative transcription factors. Microarray expression profiles of four up-regulated ApiAP2s (TA11145, TA12015, TA16485 and TA13515) were compared to the expression patterns of datasets of genes (selected based on motif enrichment analysis (see Sections 3.3.6.1 and 3.3.6.2)). Temporal expression pattern comparison was performed for TA12015 with average expression pattern of the TashAT gene family, Hsp family and merozoite to piroplasm most up-regulated genes; for TA13515 with average expression pattern of merozoite to piroplasm most up-regulated genes; for TA16485 with average expression pattern of macroschizont to merozoite most down-regulated genes and the TashAT family; and for TA11145 with average expression pattern of macroschizont to merozoite most down-regulated genes containing the CACACA type motif in their upstream region, macroschizont to merozoite most up-regulated genes and merozoite to piroplasm most up-regulated genes. Expression patterns of co-expressed genes were then presented as a graph using Excel and the Pearson correlation value was calculated (using Excel Pearson function).

3.3 Results

3.3.1 DNA motifs found in upstream regions of genes co-expressed from the macroschizont to merozoite stage

A DNA motif search was performed within 5' (upstream) and 3' (downstream) intergenic regions of the 100 most up- and down-regulated genes of macroschizont to merozoite stage differentiation. Due to software limitations, intergenic regions shorter than 8 bp (TA17346, TA13925 and TA17100) and overlapping genes (TA21395 - small overlap at the 3' end with TA21390; TA15015 - small overlap at the 5' end with TA15010) were excluded from further analysis. The three motifs of the lowest E-value (most statistically significant) identified in the intergenic regions of these stage-regulated gene sets are detailed in the Appendix (3.1-3.4). Only motifs found in 5' IGR with an E-value less than 10^{-4} and a strong core motif were considered as significant. These were further investigated in terms of their genomic location and distribution in relation to the translational start site.

Analysis of 5' IGRs of the 100 most down-regulated genes from macroschizont to merozoite stage differentiation revealed the presence of 55 instances of a 20bp motif A[TA]TT[TC][TC]AGATCCCCCAT[GT][AG][AT] (E-value = 10^{-155}) with a well conserved core motif of AGATCCCCCAT present in 44 sites (Figure 3.4). All the motif hits were located on the coding strand except the motif located upstream of TA18010, an integral membrane protein, however the last two nucleotides of the core motif were different - AGATCCCCCTA. Interestingly, two C-box motif variants were found in this dataset: a 5C-box (TCCCCCAT) associated mostly with the SVSP gene family and a 4C-box (TCCCCAT) mostly associated with the TashAT family. Sequences flanking the core motif were similar between analysed regions, with a limited number of nucleotide substitutions present. All the identified sites of the C-box motif in the subset of down-regulated macroschizont to merozoite genes shown positional conservation close to the 3' end of the IGR and proximal to the ATG start codon (Figure 3.5). Additional MEME analysis of TashAT family genes confirmed specificity of the 4C-box motif (TTCCCCA[TA]CCA[GA]) for this subset of genes (E-value = 5.7×10^{-43}) in 13 sites (Figure 3.6). TA03140, TA03150, TA03155 genes had a point mutation in the core

sequence (T instead of C), insertion of an A after the 4C sequence and the sites were localised on the negative strand in comparison to the other 13 family genes. No motif was found by MEME for TA03110 which is the only up-regulated gene in this family. Motifs present on the positive strand were localised close to the translational start site (TSS).

Name	Strand	Start	p-value	Sites [2]
TA18865	+	1194	1.31e-14	TAAACTCTTA ATTTCCAGATCCCCATGGA TCAAGCCTAT
TA18950	+	225	1.31e-14	TATCATATTA ATTTCCAGATCCCCATGGA TCATATCTCT
TA17485	+	1323	1.31e-14	GGAAATCTTA ATTTCCAGATCCCCATGGA TCAAATCTCT
TA16030	+	244	1.31e-14	GTAGTACAAA ATTTCCAGATCCCCATGGA ATAAAATTCT
TA05555	+	220	3.49e-13	TATCATATTA ATTTCCAGATCCCCATGGT TAAATGATAT
TA05560	+	250	3.49e-13	TATCATATTA ATTTCCAGATCCCCATGGT TAAATGATAT
TA17120	+	561	1.03e-12	GAATTACTAT ATTTTCAGATCCCCATGAA GAATAACATAA
TA09865	+	759	1.49e-12	TAAATTCTTA ATTTTTAGATCCCCATGGA TTAAAGTTAT
TA16025	+	343	2.95e-12	TTTATTTTTA ATTTTCAGATCCCCATGGG TCAATGTCAA
TA17535	+	225	5.16e-12	ACTGTAAATA ATTTTGAGATCCCCATGAA TAGAAATTTA
TA17540	+	260	5.16e-12	ACTGTAAATA ATTTTGAGATCCCCATGAA TAGAAATTTA
TA19005	+	302	6.01e-12	TATAATTATT AATACCAGATCCCCATGGA TTAATTATAA
TA17550	+	284	7.20e-12	ATATTGATAT ATTTTCAGATCCCCATGAA ATATACATAA
TA17125	+	220	9.23e-12	ATATTTTFTA ATTTTTAGATCCCCATGAA TTGATCTTTT
TA09800	+	204	1.63e-11	CAATTATAAT ATTTTCAGATCCCCATTAA TAAATTAATG
TA09430	+	201	1.63e-11	CAATTATAAT ATTTTCAGATCCCCATTAA TAAATTAATG
TA09790	+	118	1.63e-11	CAATTATAAT ATTTTCAGATCCCCATTAA TAAATTAATG
TA09805	+	222	1.63e-11	CAATTATAAT ATTTTCAGATCCCCATTAA TAAATTAATG
TA19060	+	414	2.26e-11	TAAGACTCTTA ATTTTAAGATCCCCATGGA TCATATCTCT
TA18890	+	193	3.52e-11	ACAATTTTFTA ATTTCCAGATCCCCATTAT TATATTATTA
TA05570	+	359	4.69e-11	ATAATACTTA ATTTCCAGATCCCCTATGAA TAATAATAAA
TA05545	+	203	7.71e-11	ATATTAGTAA TATTTCCAGATCCCCATGAA TAATAATAAA
TA05575	+	211	7.71e-11	ATATTAGTAA TATTTCCAGATCCCCATGAA TAATAATAAA
TA16035	+	367	1.40e-10	ATAATTTGTA TTTTTGAGATCCCCATTGA AATGTATACA
TA05580	+	245	1.40e-10	AATTTTATTG ATTTAAAGATCCCCATGGA ATAA
TA05540	+	98	2.04e-10	CAGTTTTTTG ATTTTAAGATCCCCATGGA ATAAAATTTA
TA11385	+	411	3.30e-10	AAACTATAAT ATTTGGAGATCCCCATTAT TATATTATTA
TA11410	+	250	3.30e-10	ATACTATAAT ATTTGGAGATCCCCATTAT TATATTATTA
TA09795	+	228	6.98e-10	GAAATCTTTA ATTTATCAGATCCCCATCTA ATTAATAATTA
TA05565	+	124	7.90e-10	ATATTAGTAA TATTTTAGATCCCCATTGA ATTAAGTTGA
TA11390	+	263	8.86e-10	TGGAAGAATT CATATTAGATCCCCATGGA TCATAACTCT
TA11400	+	688	2.43e-09	ACTATTCTTA ATTTTTAGATCCCCATGAT TAAAATTGTT
TA09435	+	660	7.86e-09	TAAATTCTTA ATTTCAAGATCCCCATATA ATTAAGTTGA
TA18010	-	430	4.93e-08	AACAAGATTA ATTTATAGATCCCCTACTCT TAAGGTAGGC
TA15710	+	496	2.37e-07	ACATCTAATA TTTATTTGATACCCCATCCA TTGATCAATG
IA15705	+	325	2.07e-06	TCATGACAAAC AATCGGTTTATACCCCATCCA CTTGATCTAC
TA17545	+	496	2.39e-06	TTTTGTAATT TATTATAGATCCCCTTTAAA TAAAATTTAT
TA03145	+	306	2.56e-06	AATGGAATCT ATTGTTTATTTCCCATCCA GATCTATAAT
TA10735	+	550	2.56e-06	GTTGAAAACC ATTTCCAGATCTGCCCTCTTA ATTATTGATA
TA20095	+	581	2.56e-06	GAATCTAAA ATCTCTTAGTTCCCATCCA GTTGCTTTAA

Figure 3.4. ATCCCCAT motif found in the upstream region of the 100 most down-regulated genes during differentiation of the macroschizont to merozoite stage in *T. annulata*.

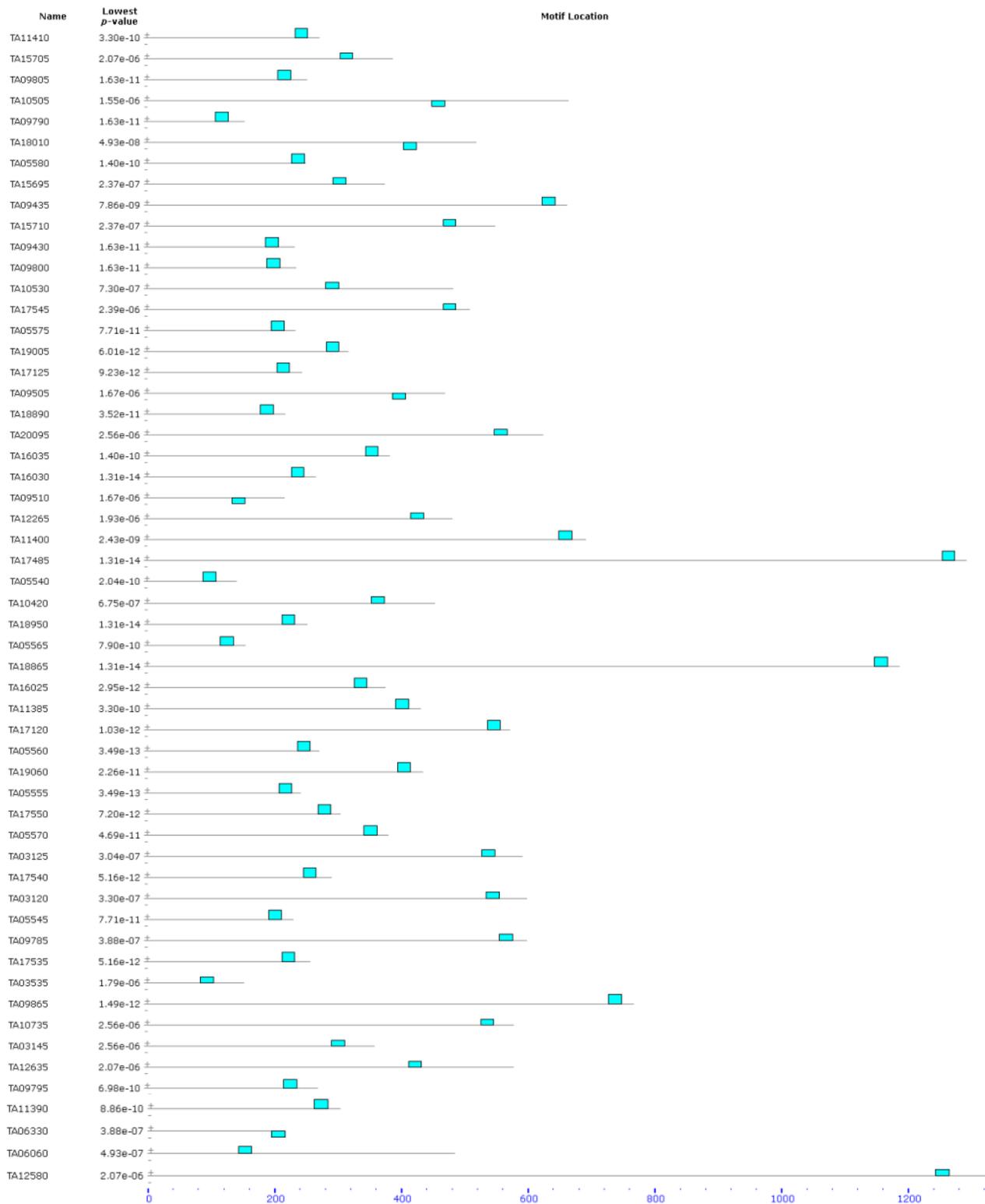


Figure 3.5. Distribution of the ATCCCCAT motif within intergenic regions of down-regulated genes. The predicted translational start site is at the 3' end (right side) of the intergenic region.

Name	Strand	Start	<i>p</i> -value	Sites [?]
TA03115	+	661	1.21e-09	TATTATTTAA TTCCCATCCAG ATCTATCAGT
TA03125	+	561	1.21e-09	AAATTGTTAA TTCCCATCCAG ATCTATAAAA
TA03130	+	561	1.21e-09	AAATTGTTAA TTCCCATCCAG ATCTATAAAA
TA03135	+	555	1.21e-09	TTTTAATTGG TTCCCATCCAG ATCTATAGAA
TA03145	+	315	1.21e-09	TATTGTTTAT TTCCCATCCAG ATCTATAAAT
TA03160	+	493	1.21e-09	TATTGTTTAT TTCCCATCCAG ATCTATAAAT
TA20082	+	407	1.21e-09	AATATTTTAT TTCCCATCCAG ATCTAGCAAT
TA20085	+	407	1.21e-09	AATATTTTAT TTCCCATCCAG ATCTAGCAAT
TA20095	+	590	1.21e-09	AATCTCTTAG TTCCCATCCAG TTGCTTTAAT
TA03120	+	568	7.10e-09	AATTCGTTAA TTCCCATCCAA ATAATTAATA
TA03165	+	439	7.10e-09	TTTTGTTTAT TTCCCATCCAA ATCTATAAAA
TA20083	+	381	2.66e-07	TATTAATTAT TTCCCATCAAT TATAGGAACA
TA20090	+	381	2.66e-07	TATTAATTAT TTCCCATCAAT TATAGGAACA
TA03140	-	59	1.40e-06	TCATAAATTA TTCTCCAATCAA TCTATCATTA
TA03150	-	173	1.40e-06	TCATAAATTA TTCTCCAATCAA TCTATCATTA
TA03155	-	73	1.40e-06	TCATAAATTA TTCTCCAATCAA TCTATCATTA
TA03110	-	83	4.60e-06	AATTTTTTAA TTACACAACCAA CCTCATTTCA

Figure 3.6. Motif 'TCCCAT' (4C-box) found upstream of TashAT family genes.

Analysis of 5' IGR of 100 most up-regulated genes from macroschizont and merozoite stage revealed presence of 50 sites of a 14 bp motif [AG]AATGTGTAA[AG][GT][TAG][AT] (E-value = 1.3×10^{-9}) with conserved core motif of AATGTGTAA (Figure 3.7). This motif was found on positive and negative strands and was neither associated with a particular gene family nor positionally associated with the ATG start codon (Figure 3.8). Analysis of the 5' intergenic regions of 100 *T. annulata* constitutively expressed genes during macroschizont to merozoite differentiation also revealed the presence of this motif in 98 sites (E-value = 2.3×10^{-15}), implying a wide-genome distribution of this motif (Appendix, 3.7).

An additional MEME analysis of intergenic regions of ApiAP2 and other transcription factors (see Chapter 2) was performed to look for potential shared motifs, but did not reveal any statistically significant results (see Appendix, 3.5-3.6). This is likely due to the small number of analysed sequences. It is also possible that these putative transcription factors do not share common DNA motifs regulating their expression, with each gene regulated separately at the transcriptional level by incorporating a unique arrangement of a number of distinct binding site motifs. This would be unlikely to be detected by the repeat pattern software employed in this study.

Name	Strand	Start	p-value	Sites
TA16685	-	513	2.82e-08	CGCGTGATTT GAATGTGTAAGGA TTTCCATATAA
TA14680	-	353	8.74e-08	GATAAATTAG AAATGTGTAAGGAA TTTAGGTCAA
TA14665	+	773	1.52e-07	TGATGTGTAA AAATGTGTAAGTGA CCCAATAAAA
TA05340	+	291	2.11e-07	AATTCGGAT AAATGTGTAAGTA CGATTACGAA
TA12015	+	90	4.36e-07	AAATCGTCTA AAATGTGTAATGA AGTTCAAAGT
TA17705	-	588	5.00e-07	TTACTAACAT GAATGTGTAAGAG AAAAACATAG
TA13045	+	116	8.64e-07	AATATTAGAG AAATGTGTAGAGAA GGATGCACCT
TA21080	+	960	8.64e-07	TTTTTGGCTG AAATGTGTAAGCAA ATGGTTATAT
TA19675	+	207	1.06e-06	AATTATAAAT GAATGTGTAAGTGT TTGGAATAAA
TA20555	-	602	1.06e-06	AGTTACATAA AAATGTGTAAGTAA ATTCATATGTT
TA03260	-	137	1.06e-06	AAAAAAACAA AAATGTGTAAGTAA ATTAACATAG
TA13515	+	56	1.25e-06	TATAGAACCA AAATGTGTGAGGGT AAAAATTACC
TA21050	-	347	1.46e-06	GATTTGGGAT AAATGTGTAATGT TTAATTAGT
TA03540	-	484	1.46e-06	ATATTAGATA GAATGTGTATGGTA TCAGGTCTAA
TA04355	+	981	1.46e-06	ACTAATATTT AAATGTGTAACCTA GTGACAACAG
TA13890	+	235	1.46e-06	ATTTTAGATA GAATGTGTATGGTA TCAGGTAGTT
TA11285	+	59	2.08e-06	AGTAGTAATT AAATGTGTAATGA ATTTATTAGT
TA05495	-	309	2.08e-06	AATAAAATTA AAATGTGTAATTA TTAACATTAA
TA05870	+	270	2.08e-06	AAAAATFACT AAATGTGTAATTA AGAATGTATA
TA19575	-	411	2.44e-06	GAGAAAACCT AAATGTGTATAGTA AAAACTGTAA
TA11455	+	234	2.44e-06	ATAAAAAACC AAATGTGTAATTTG GAATGGAATC
TA13345	+	220	2.79e-06	GTTTTAATTA AAATGTGTGAAGTA TTTTAAAGCTA
TA19275	-	598	2.79e-06	CTGATGTAAT GAATGTGTAAGA ATTATAGTGA
TA20985	+	148	3.54e-06	TAATGTATAA AGGTGTGTAAGGAG ATCTTCTGCG
TA16485	-	401	3.98e-06	ACGGGAGTGG AAATGTGTAGGAG TAATTGAGGT
TA05760	+	316	3.98e-06	ATATAAAGT AAATGTGTAGACAA TGAATACGA
TA04565	-	221	4.55e-06	ATACACCGAT GGATGTGTAAGAGA TATGCAAATT
TA04105	-	24	4.55e-06	AATATTGGAT AAATGTGTAATTTT TAAGAAGGAT
TA16375	+	342	4.55e-06	AGTTTATTAG AGATGTGTAATGT GATGGAACGT
TA07985	-	199	5.13e-06	AATTCCACTT AGATGTGTAACAA ATCATATAGT
TA20020	-	542	5.13e-06	GCTATTTTCC AGATGTGTAATTA AAACACAATT
TA19610	+	79	5.75e-06	GTAATTTTTA AAATGTGTAGCCGT TTAGTGACTT
TA07630	-	136	7.82e-06	CATATATAAA TAATGTGTAAGGAT AAAATCACAT
TA13825	-	351	7.82e-06	TTTAAAGATT AAATGTGTAGACTT TTTTATATGA
TA16660	-	276	1.14e-05	CAAAAAGTTA AAATGTGTAACAAA ATTTAGTCAA
TA07920	-	452	1.24e-05	AAAATTTACC AAATGTGTATGTAT TTTACAGTTA
TA03755	-	216	1.34e-05	TGCCTATCTT TAATGTGTGAGGTA GAATCTCGTG
TA19975	-	606	1.34e-05	ACAATAGACT AGATGTGTAATA AATTTTACTT
TA07435	+	101	1.45e-05	AGCTGTTAAT AAGTGTCTAAGGTG CAGTTTAGAC
TA17325	-	282	1.57e-05	TATTAGATTA AAATGTGTATCCAG AAACGGATA
TA19445	+	588	1.70e-05	GAAAGTGGTT GAATGTGTATATTT TTGAGTAAAA
TA03480	-	281	1.83e-05	TCTAGACCTG GAGTGTGTAACCTT AACCAAATTT
TA07585	+	968	1.96e-05	TTTAGAACAA AAATGTCTAATTA TCGGTAATGA
TA10690	+	102	2.11e-05	AATTATTCGT TAATGTGTACATGA TTTTACAAGG
TA13535	+	102	2.24e-05	TTATTCTAAA GAATGTCTACAGGT TATATTCGT
TA13530	-	163	2.24e-05	TTATTCTAAA GAATGTCTACAGGT TATATTCGT
TA14835	-	317	2.53e-05	AATTA AAAAC AAATGTGTATCAGT ATATAACATG
TA19075	-	44	2.53e-05	CAGCCTAATA TAGTGTGTATGGAA TGTTAACGGC
TA08360	+	206	3.75e-05	GTGTATGAAA AGGTGTGTGGCAG AATGAAGACG
TA11905	-	200	5.32e-05	GATTTGATAA TAATGTGTAATGTG TAATCTATAC

Figure 3.7. TGTGT motif identified in the 5' intergenic regions of the 100 most up-regulated genes during macroschizont to merozoite stage differentiation in *T. annulata*.

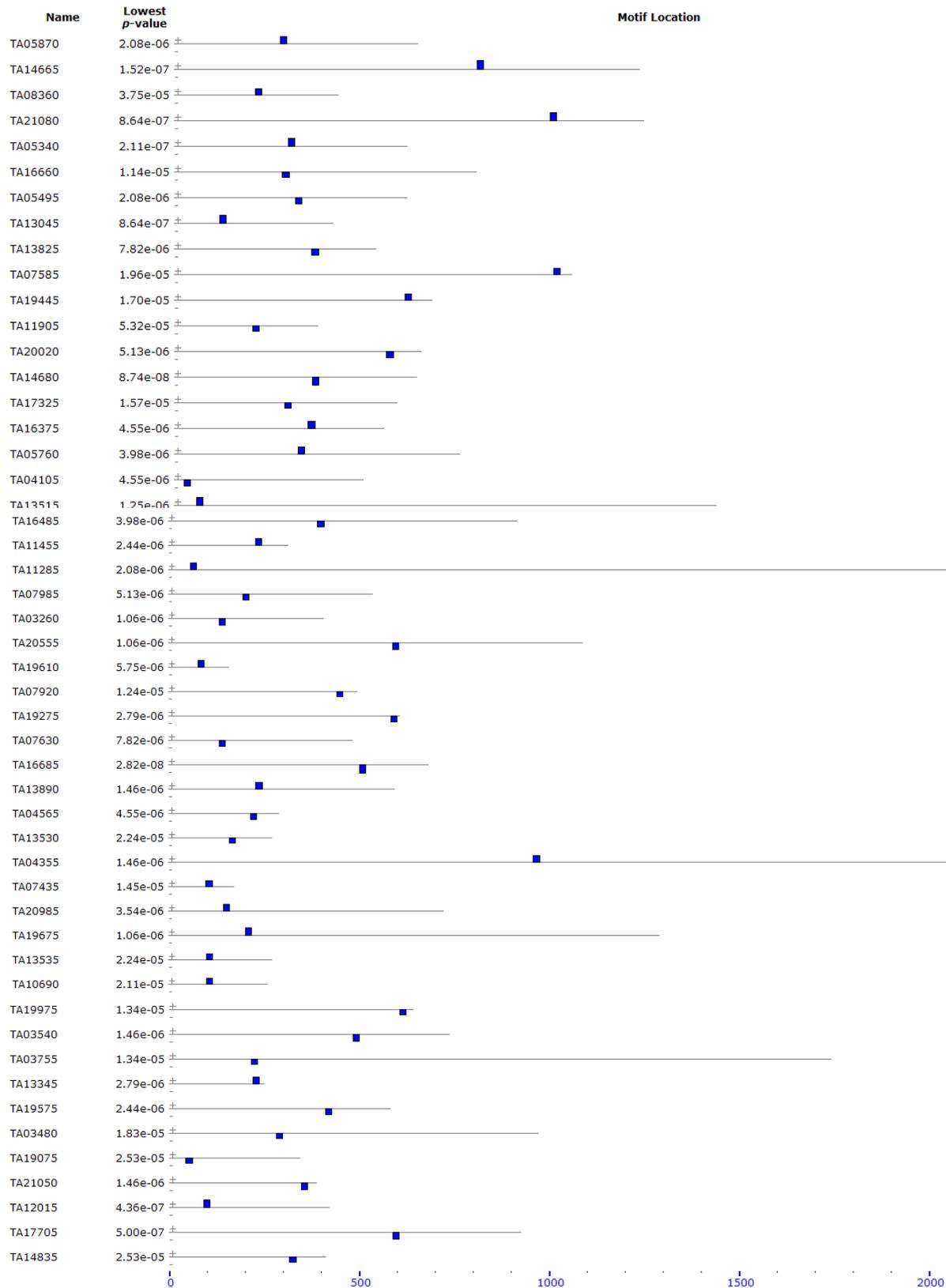


Figure 3.8. Distribution of the TGTGT motif upstream of up-regulated genes. The predicted translational start site is at the 3' (right hand end) of the intergenic region.

3.3.2 Cross-species analysis for the TashAT cluster

The *T. annulata* TashAT gene family consists of 16 down-regulated and only one up-regulated gene, TA03110. MEME analysis of upstream regions revealed a 'TCCCCAT' (4C-box) motif present on the positive strand close to the TSS. The TashAT family is conserved between *T. annulata* and *T. parva* and encodes proteins predicted to control host cell gene expression. Although a number of direct orthologues have been identified across the two families it is clear that there has been species-specific expansion and divergence. Thus, towards the centre of the locus in *T. annulata*, genes without direct orthologues in *T. parva* are evident, and DNA binding domains (AT hooks) present in TashAT cluster genes are not present in the TphN family of *T. parva* (Shiels et al., 2006). The availability of a family of genes that have diverged between closely related species may provide further evidence that motifs, and the proteins that bind to them, are functionally conserved.

Cross-species MEME analysis of the intergenic regions of *T. annulata* and *T. parva* TashAT cluster genes has been then performed and an enlarged 4C-box motif, 'TTCCCCATCC', identified that is present in the majority of genes in both species (Figure 3.9). Strong positional conservation of this motif, located close to the predicted translational start site was also observed (Figure 3.10). A second motif with an 'AGGGTA' core was also identified in a smaller number of genes across the two species (Figure 3.11). This motif also showed a strong positional conservation, being close to the predicted translational start site (Figure 3.12) but always upstream of the C-box rich motif.

NAME	STRAND	START	P-VALUE	SITES		
up_0612	+	391	1.62e-19	GATAAAGTTA TCAATTTATTC	CAGATCTATGAAAAATATTATAAGATTGAAATA	ATATTCCTTTC
up_0607	+	360	2.04e-19	GATAAAGTTA TCAATTTATTC	CAGATCTATGAAAAATATTATAAGATTGACATG	GTAAACGTGA
TA_up_Tash_like_b	+	654	2.56e-19	ATCTAAATTA TATATTTAAAT	CAGATCTATCAGTAATTTTATAAGATTTAGTAA	TTT
up_0618	+	361	1.39e-18	TTAATCTAAA AATGTTAATTC	CAGATCTAGCAATGATCCTGTAAGATTTAATAG	TACGAGAATG
up_0620	+	530	2.06e-18	ACTTAATGTA TTAGTTAAAT	CAGATCTATCGATAAAATTAAGATATACAAA	ATACAAATGA
up_0602	+	429	5.62e-17	AATCTAAAAA TTGCTTAAAT	CAGATCCGTCAATAATTTCAAGTGCTTTTATA	AACGAGGTGG
TA_up_TashAT3	+	400	6.57e-17	TTAATCTAAA TATTTTATTC	CAGATCTAGCAATTACTGTGTTATTTTAAATAG	TAGTATAAGC
TA_up_TashAT1	+	400	6.57e-17	TTAATCTAAA TATTTTATTC	CAGATCTAGCAATTACTGTGTTATTTTAAATAG	TAGTATAAGC
up_0614	+	462	5.01e-16	ATGGAATATT TTAATTTGAT	CAAAATTTAGCATTGATTCGTAAAGATTTATAG	TACAGAAATG
TA_up_Tash_like_d	+	308	1.44e-15	ATGGAATCTA TTGTTTATTC	CAGATCTATAATTAATATTTATTAATAAATA	ATAGATTTAT
TA_up_2A1	+	486	1.44e-15	ATGGAATCTA TTGTTTATTC	CAGATCTATAATTAATATTTATTAATAAATA	ATAGATTTAT
up_0611	+	304	1.86e-15	TATCTAATAA TCTTTCAAT	CAGATCTATAAAAAATATTATAAGATTGAAATA	TATTCCTTTA
up_0610	+	393	1.86e-15	TATCTAATAA TCTTTCAAT	CAGATCTATAAAAAATATTATAAGATTGAAATA	TATTCCTTTA
TA_up_SuAT1	+	548	3.45e-15	ATTTATTATT TTAATTTGGT	CAGATCTATAGAAAAATTTTGTGCTTAATATA	GCACGGAAAT
TA_up_Tash_like_f	+	432	3.90e-15	TTAAACATTT TTGTTTATTC	CAAAATCTAAAACTATATTAATAAATAATAG	ATTTATGATT
up_0606	+	447	4.96e-15	ATGGAATATT TTAATTTAAAT	CAAAATTTTGAATAATATTTATAAAATTTAATGG	TACGGGAAAG
TA_up_Tash1	+	554	4.96e-15	TTAATCTAAA ATTGTTTAAAT	CAGATCTATAAAACCTATATTTATTAATAAATA	AAATATGATT
TA_up_SuAT2	+	554	4.96e-15	TTAATCTAAA ATTGTTTAAAT	CAGATCTATAAAACCTATATTTATTAATAAATA	AAATATGATT
up_0617	+	387	3.34e-14	AATTTAATCT ATATTTAAAT	CAGATCTATAAAAAATATTATAAGATTGAAATA	TATTCCTTTA
up_0604	+	388	8.45e-14	GGAAATCTAA AATTTTAAAT	CAGATCTAACAATGATCCTGTAAGATTTAATA	TAAGGGAAATG
up_0613	+	369	2.34e-12	GATTAACGGT TTTTATAAT	GTGACCCAGTGAATAACAACATAAAATTTTAAAT	AGTTAAAAAT
up_0609	+	380	2.34e-12	GATTAACGGT TTTTATAAT	GTGACCCAGTGAATAACAACATAAAATTTTAAAT	AGTTAAAAAT
up_0605	+	348	2.34e-12	ATGAAGCTTT TTTTCTAAAT	AAATCTTATCAATAATATTTTGTATTTTAAATA	CTAATAACTT
up_0616	+	329	7.98e-12	ATGGAATATT TTAATTTGAT	TTGATCAATGATTTTTTTTGTATTAATAAGTATG	GGAAATATGA
up_0615	+	329	7.98e-12	ATGGAATATT TTAATTTGAT	TTGATCAATGATTTTTTTTGTATTAATAAGTATG	GGAAATATGA
TA_up_TashAT2	+	583	1.23e-11	AATCTAAAAA TCTCTTAAAT	CAGTGTGCTTAAATAAACCATTGTTGTTTAAATG	CAGATAAGTT
TA_up_Tash_like_c	+	561	3.40e-11	AATCCATAAA TTGCTTAAAT	CAAAATTAATAAATAATCCAAATCCATAAATAAG	ATTTATGATT
up_0603	+	353	1.73e-10	ATATAAATCTG TTAATTTAAAT	CAAAATTCATAATCATTTATTAATTCAAAGACTA	GAGTTGAATC
TA_up_TashHN2	+	374	8.23e-10	ATATAAATCTA TTAATTTAAAT	AAATTAAGGAAACAATTTTATTAAGTTAAATATAT	AGGATTGAGT
TA_up_TashHN1	+	374	8.23e-10	ATATAAATCTA TTAATTTAAAT	AAATTAAGGAAACAATTTTATTAAGTTAAATATAT	AGGATTGAGT

Figure 3.9. Cross-species MEME analysis of the intergenic regions of *T. annulata* and *T. parva* TashAT cluster genes. The 4C-box motif was identified within this region of similarity. Discovered sites were aligned with each other. Each site is identified by the sequences name (TA - *T. annulata*, remainder - *T. parva*), the strand (positive) and the position in the strand where the site begins. Sites are listed in order of increasing statistical significance (p-value).

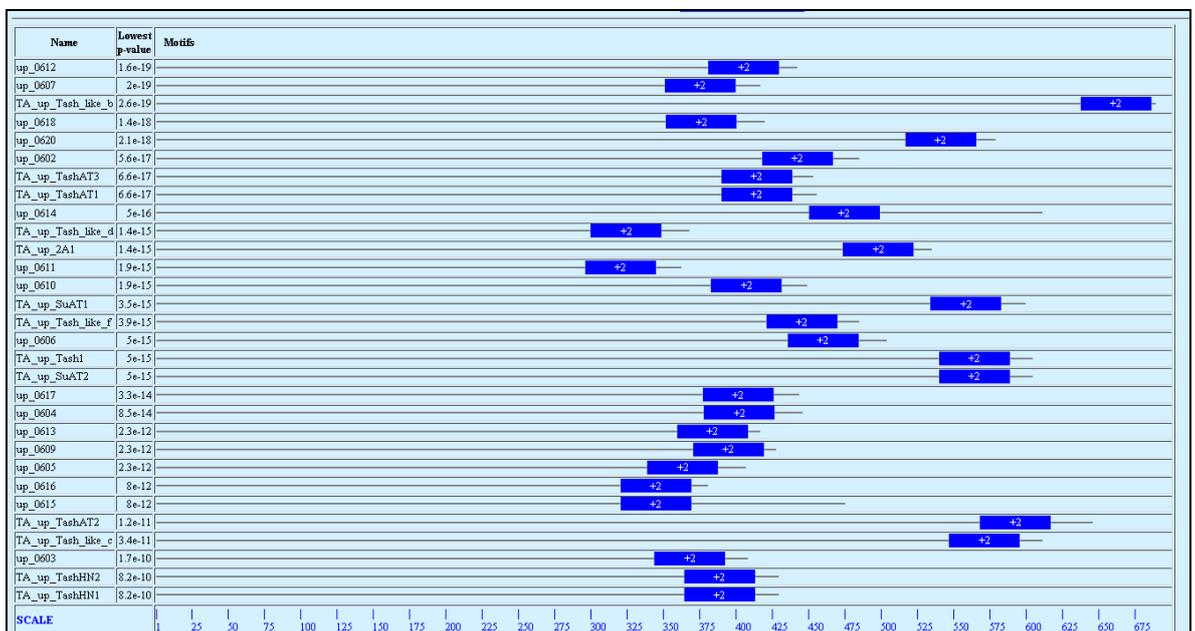


Figure 3.10. Distribution of the 4C-box motif within upstream regions of down-regulated TashAT/TPHN family genes. The predicted translational start site is at the 3' (right end) of the intergenic region. Strong positional conservation of the motif is evident.

NAME	STRAND	START	P-VALUE	SITES
up_0616	+	216	3.02e-23	TATTAAAAAT AGCGTGTATGTRATATGARRAGGGGTAATATCGCTACTGATTTTGATA TAAAATATAT
up_0615	+	216	3.02e-23	TATTAAAAAT AGCGTGTATGTRATATGARRAGGGGTAATATCGCTACTGATTTTGATA TAAAATATAT
up_0606	+	333	6.34e-21	TATTAAAAAT AGTGTTTATGTRATATGARRAGGGGTAATATCGCTACTGATTTTGATA TAAAATATAT
TA_up_Tash1	+	446	7.08e-20	TTTAAATCCA AAAATRRRRGARRATATATARRAGGGGTAATATATTTTGGRATARTCGATG ATATATGGTT
TA_up_SuAT2	+	446	7.08e-20	TTTAAATCCA AAAATRRRRGARRATATATARRAGGGGTAATATATTTTGGRATARTCGATG ATATATGGTT
TA_up_SuAT3	+	331	7.08e-20	TTTAAATCCA AAAATRRRRGARRATATATARRAGGGGTAATATATTTTGGRATARTCGATG ATATATGGTT
TA_up_Tash_like_e	+	445	7.08e-20	TTTAAATCCA AAAATRRRRGARRATATATARRAGGGGTAATATATTTTGGRATARTCGATG ATATATGGTT
up_0608	+	208	1.05e-19	TATTATAAAT AGTGTGTATGARRATCGARRAGGGGTAATATATCTTGTATARTGATG ATTAATATAT
TA_up_S2	+	235	1.76e-17	TGTATTTAAT ACTTTATATGARRATATARRAGGGGTAATACACCTTCTAATGATTGATA ATTGATATAT
up_0611	+	205	3.16e-16	ATGTATAAAT GCTATATACTARRATGARRAGGGGTAATARRGTTGACGACTCTTGTTA TACAATGTGT
up_0610	+	294	3.16e-16	ATGTATAAAT GCTATATACTARRATGARRAGGGGTAATARRGTTGACGACTCTTGTTA TACAATGTGT
TA_up_2A1	+	367	1.20e-15	TTATTAAAAAT ACTTTATATGARRATATARRAGTGTGATAGRGATAGTTATRTTGATA CATCATATAT
up_0617	+	280	3.80e-15	ATTGAGACAA AACGTRRRRTARRRRATARRAGGGGTAATATATTTTGGRARRATAGATG ATATATGGTT
TA_up_Tash_like_f	+	330	4.71e-14	TTTAAATCCA AAAATRRRRGARRATATATARRAGGGGTAATATATTTTGGRARRATAGATG ATATATGGTT
TA_up_SuAT1	+	419	8.11e-14	CTACTAATAT TGTARRATATGATATARRAGGGGTAATACGCTTGTARRARTTGATA CTTAATATAT

Figure 3.11. Cross-species MEME analysis of the upstream regions of *T. annulata* and *T. parva* TashAT cluster genes. A second conserved, TAGGGTA, motif was identified. Discovered sites were aligned with each other. Each site is identified by the sequences name (TA - *T. annulata*, remainder - *T. parva*), the strand (positive) and the position in the strand where the site begins. Sites are listed in order of increasing statistical significance (p-value).

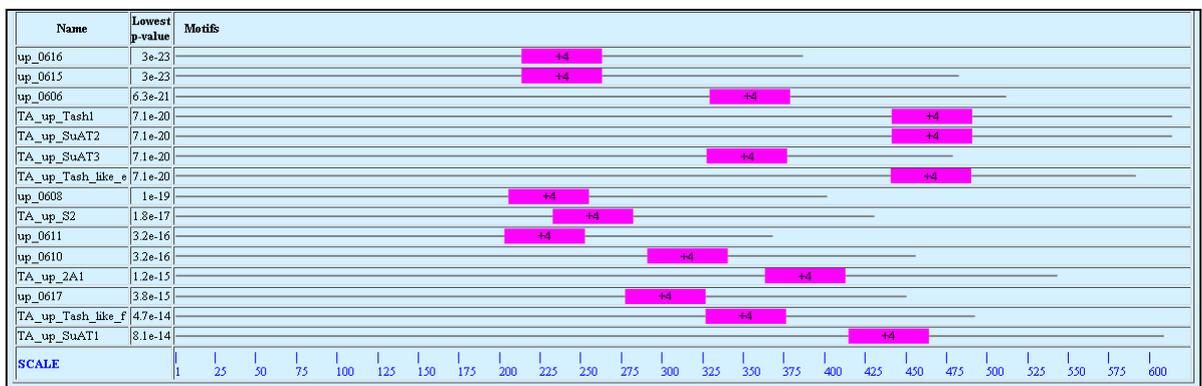


Figure 3.12. Distribution of the GGGTA motif within intergenic regions of down-regulated TashAT/TpHN family genes. The predicted translational start site is at the 3' (right hand end) of the intergenic region.

3.3.2.1 MEME analysis of the intergenic region of TA03110 (Tash-a) across *T. annulata*, *T. parva* and *T. orientalis*

To determine if the 4C-box motif is associated with down-regulation of genes in the TashAT cluster, analysis of the intergenic region of the gene encoding Tash-a (TA03110) was performed. Tash-a is the only gene in the cluster that is conserved in the non-transforming *Theileria* species, *T. orientalis*, and is up-regulated during merogony, based on microarray data and IFAT (Hyashida et al., 2013). The 4C-box motif MEME alignment generated for the TashAT cluster did not include any motifs in the intergenic region of the TA03110 gene, indicating that this region was divergent relative to the rest of the gene family.

Motif searching in the upstream region of the TA03110 (Tash-a) gene and its orthologues in *T. orientalis* and *T. parva* was carried out using MEME with a motif length of between 6 and 14bp and zoops (zero or one occurrence per sequence) options (Figure 3.13). A 4C-box motif was found on the coding strand upstream of the *T. orientalis* gene and on the non-coding strand upstream of *T. parva* gene, however the flanking nucleotides differed from each other and the full length *T. annulata* C-box motif in Figure 3.9. The sequence upstream of TA03110 was also visually analysed and a 'TCCCCAT' motif was identified on the coding strand (Figure 3.15). This motif shares a similar position to the previously identified 4C-box motif (Figure 3.14), close to the predicted translational start site of TA03110. The lack of recognition of these motifs in TA03110 by MEME is most likely due to divergence from the full length C-box motif conserved in the IGRs of other TashAT/TpHN family genes (particularly on the 3' side of the core 4C-box) and the more stringent parameters of the MEME search. These results suggest that if the motif independently promotes differential expression between the up-regulated TA03110 and down-regulated members of the cluster, it is due to this divergence.

Name	Strand	Start	p-value	Sites ?
T.orientalis_TA03110_upstream_IGR	+	1484	2.33e-06	GCCTAAGTTT GAGTCCCC CGCCGCAACT
TP01_0621_upstream_IGR	-	160	7.03e-06	AAATATGTTA GAATCCCC AAGAAGTCAG

Figure 3.13. Cross-species MEME analysis of the intergenic regions of *T. annulata*, *T. orientalis* and *T. parva* TA03110 orthologues. A 4C-box motif was found on the coding strand of *T. orientalis* and the non-coding strand of *T. parva* (This C-box is a G-box on the coding strand).

Name	Strand	Start	p-value	Sites ?
TP01_0621_upstream_IGR	-	252	2.81e-05	TAATGCATAT GGGGATC TTTTTTACAC
T.orientalis_TA03110_upstream_IGR	+	1282	2.81e-05	AATAGTAGGG GGGGAAC GACGCGGTGA

Figure 3.14. Cross-species MEME analysis of the intergenic regions of *T. annulata*, *T. orientalis* and *T. parva* TA03110 orthologues. A G-box motif was found on the coding strand of *T. orientalis* and the non-coding strand of *T. parva* (This G-box is a C-box on the coding strand).

```
5' -ATATAAATTGAACCTAATTAATTTAAAAACAAGACTTTAAAAACAGTAATATCTAAATTAAT
GTAAAATTAAAAAATGAAATGAGGTTGGTTGTGTAATTTAAAAATTTGATGAATAGAGGTT
AATTTCCCTCAAGATTCCAACCAATTTACAACAACACATTTAAATGATAAGAGGAAAAATAC
TAGCATAATTAAGGATTAAGTTAAAAAATGTGTAAAAACGATCCCCATATGCATTAACGA
TAAAAATCGCATGTTGAAGCGTTATAATATGTTTTAACTTAATAT-3'
```

Figure 3.15. The sequence upstream of TA03110. A core 4C-box motif is highlighted in green.

3.3.3 Statistical enrichment analysis of 4C-box motif

Based on MEME analysis of stage-regulated macroschizont to merozoite genes, a C-box motif was identified as associated primarily with genes down-regulated during differentiation towards tick transmissible stages. To confirm specificity and show any statistical enrichment of this motif in *T. annulata* subsets of genes, a Motif Pattern Search was performed using PiroplasmaDB. The 4C-box motif, 'TTCCCAT', was found on the positive strand of target genes up to 200bp upstream of the start codon. Genes that possess this motif within their 5' intergenic region are presented in a Table 3.1. and include genes from TashAT family, Sfil-subtelomeric fragment related protein family members, ribosomal subunit proteins, ATP binding proteins and nucleic acid binding proteins.

Gene family	Gene with 4C-box motif in the upstream region
TashAT family	TA03115, TA03120, TA03125, TA03130, TA03135, TA03145, TA03160, TA03165, TA20082, TA20083, TA20085, TA20090, TA20095 (down-regulated genes) and TA03110 (the only up-regulated TashAT gene)
Sfil-subtelomeric fragment related protein family members	TA16050, TA16055, TA17105, TA17110, TA17115, TA17355, TA17495, TA17505, TA17565, TA02845, TA05530, TA05535, TA09450, TA09770, TA09775, TA11395, TA11400, TA11405
ribosomal subunit proteins	TA04505, TA07390, TA08205, TA10185, TA14730, TA02825, TA10305, TA14270, TA15200, TA20525
ATP binding proteins	TA10040, TA21325, TA03485, TA05325, TA10790, TA11560, TA12510, TA20945
nucleid acid binding proteins	TA03320, TA05440, TA06205, TA11880, TA14280, TA14520, TA19905, TA14970, TA20210

Table 3.1. Gene families enriched with the 4C-box motif in their upstream regions.

Of the 3,856 *T. annulata* genes screened using PiroplasmaDB, 536 4C-box motifs were found, but only 170 were placed 200bp upstream of the translational start site on the (+) coding strand of the gene. The majority of these genes were identified as down-regulated based on microarray data or expressed at the same level during the differentiation process. Some up-regulated genes (e.g. TA03110, Tash-a) were also found to have the core motif in their upstream region. Enrichment analysis for the 4C-box motif showed that within a 400bp upstream region on the positive strand 82% of TashAT genes (14 out of 17 genes) possess the motif (Table 3.2). In the group of down-regulated from macroschizont to merozoite genes, 12 out of 113 genes (11%) shared this motif in the 400bp

upstream region and showed a statistically significant enrichment relative to the rest of the dataset. Narrowing the area for potential location to 100bp proximal to the translational start site confirmed that the motif is close to this site in TashAT family genes (82%). Also, the motif was found to be statistically enriched in the region 100 bp upstream of the genes which are most down-regulated between the macroschizont to merozoite. The motif was also enriched in the region 400bp upstream Heat Shock Protein encoding genes on the negative strand, which is a G-box on the positive strand.

TCCCCAT motif	TashAT genes	non-TashAT genes	macro-mero most down-regulated genes	non-macro-mero most down-regulated genes	macro-mero most up-regulated genes	non macro-mero most up-regulated genes	mero-piro most up-regulated genes	non-mero-piro most up-regulated genes	mero-piro most down-regulated genes	non-mero-piro most down-regulated genes	SVSP genes	non SVSP genes	HSP genes	non HSP genes
400 bp upstream positive strand	82.35% (14/17) P<0.0001	5.49% (206/3751)	10.62% (12/113) P=0.0217	5.7% (208/3665)	8.1% (12/148) P=0.2121	5.74% (208/3620)	12.5% (3/24) P=0.1619	5.78% (217/3743)	10% (2/20) P=0.1077	5.8% (218/3747)	2% (1/48) P=0.5265	5.9% (219/3719)	0% (0/34) P=0.2619	5.89% (220/3733)
400 bp upstream negative strand	0% (0/17) P=1	4.08% (153/3751)	0.88% (1/113) P=0.0896	4.16% (152/3655)	2.02% (3/148) P=0.2848	4.14% (150/3620)	12.5% (3/24) P=0.0713	4% (150/3743)	5% (1/20) P=0.5646	4% (152/3747)	0% (0/48) P=0.2637	4.11% (153/3719)	8.8% (3/34) P=0.0471	4% (150/3733)
100 bp upstream positive strand	82.35% (14/17) P<0.0001	2.38% (87/3655)	7.08% (8/113) P=0.0071	2.38% (87/3568)	5.4% (8/148) P=0.1317	2.4% (87/3620)	12.5% (3/24) P=0.0027	2.45% (92/3743)	0% (0/20) P=1	2.53% (95/3747)	2% (1/48) P=1	2.53% (94/3719)	0% (0/34) P=1	2.54% (95/3733)

Table 3.2. Statistical enrichment analysis of 'TCCCCAT' motif of *T. annulata* intergenic regions within 400 bp upstream of the translation start site. Statistically significant data (P value < 0.05, Fisher's Exact Test) are shown in yellow.

3.3.4 Enrichment analysis of 5C-box motif

The 5C-box motif, 'TCCCCCAT', was identified on the positive strand within 200bp upstream of the start codon using the DNA Motif Pattern Search of PiroplasmaDB. Of 3,856 screened *T. annulata* genes, this motif was identified in 98 segments but only 52 were placed within 200bp upstream of the predicted translational start site on the coding (+) strand of the gene. Annotated genes containing the 5C-box motif in their upstream regions are presented in the Table 3.3. and, in addition to hypothetical genes, include genes encoding: SVSP (subtelomeric variable secreted proteins) family, a u3 small nuclear ribonucleoprotein, a u2 snRNP auxiliary factor, the 60S ribosomal protein l15, a serine/threonine-protein kinase, a u5 snRNP-specific subunit, a micronemerohoptry antigen and a vacuolar ATP synthetase subunit beta. Only two of these genes, TA13535 and TA11485 (hypothetical proteins), were identified as up-regulated during differentiation based on microarray data. The motif was also found on the non-coding strand (-) strand within 200bp upstream of 12 genes. Of these, only the phosphate transporter gene (TA13530) was identified as being up-regulated during the macroschizont to merozoite differentiation process.

Annotation	Genes containing 5C-box motif in upstream region
SVSP (subtelomeric variable secreted proteins) family	TA19060, TA190005, TA18950, TA18890, TA18885, TA18865, TA18860, TA16040, TA16035, TA16030, TA16025, TA11410, TA11390, TA11385, TA09430, TA09435, TA17485, TA17540, TA17535, TA02740, TA05580, TA05575, TA05565, TA05560, TA05555, TA05550, TA05545, TA05540, TA17120, TA17125, TA20460
u3 small nuclear ribonucleoprotein	TA19780
u2 snRNP auxiliary factor	TA03790
60S ribosomal protein l15	TA18090
serine/threonine-protein kinase	TA18470
u5 snRNP-specific subunit	TA09305
<i>T. parva</i> microneme-rhoptry antigen	TA08425
vacuolar ATP sythetase subunit beta	TA08450

Table 3.3. List of genes with a 5C-box motif in their upstream region

For motif enrichment analysis, a p-value of 0.05 or less was taken to indicate statistical significance. Analysis of the 5C-box motif located within the 400bp upstream region on the positive strand showed enrichment (27.43 %) in IGRs of macroschizont to merozoite down-regulated genes (Table 3.4). The motif was also found to be strongly associated with SVSP genes and showed positional conservation; 70.83 % SVSP IGRs possess this motif within 100bp upstream of the ATG codon.

TCCCCCAT motif	TashAT genes	non-TashAT genes	macro-mero most down-regulated genes	non-macro-mero most down-regulated genes	macro-mero most up-regulated genes	non-macro-mero most up-regulated genes	mero-mero most up-regulated genes	non-mero-mero most up-regulated genes	mero-mero most down-regulated genes	non-mero-mero most down-regulated genes	SVSP genes	non-SVSP genes	HSP genes	non-HSP genes
400 bp upstream positive strand	0% (0/17) P=1	1.413% (53/3751)	27.43% (31/113) P<0.0001	0.6% (22/3655)	0.67% (1/148) P=0.7227	1.43% (52/3620)	0% (0/24) P=1	1.41% (53/3743)	5% (1/20) P=0.24	1.39% (52/3747)	70.83% (34/48) P<0.0001	0.51% (19/3719)	0% (0/34) P<0.0001	1.42% (53/3733)
400 bp upstream negative strand	0% (0/17) P=1	0.48% (18/3751)	0.88% (1/113) P=0.0999	0.46% (17/3665)	0.67% (1/148) P=0.5147	0.47% (17/3620)	0% (0/24) P=1	0.48% (18/3743)	0% (0/20) P=1	0.48% (18/3747)	0% (0/48) P=1	0.48% (18/3719)	0% (0/34) P=1	0.48% (18/3733)
100 bp upstream positive strand	0% (0/17) P=1	1.09% (41/3751)	26.55% (30/113) P<0.0001	0.3% (11/3655)	0% (0/148) P=0.4083	1.13% (41/3620)	0% (0/24) P=1	1.09% (41/3743)	0% (0/20) P=1	1.09% (41/3747)	70.83% (34/48) P<0.0001	0.19% (7/3719)	0% (0/34) P=1	1.1% (41/3733)

Table 3.4. Statistical enrichment analysis of ‘TCCCCCAT’ motif within *T. annulata* intergenic regions (400 bp upstream of the translation start site). Statistically significant data (P value < 0.05, Fisher’s Exact Test) are shown in yellow.

3.3.5 ApiAP2 protein conservation across *Theileria* and *Plasmodium* species

Based on previous studies, ApiAP2 domains are known to be conserved across Apicomplexan species and genera. T-coffee analysis of all *T. annulata* ApiAP2 domains confirmed strong conservation of secondary structure as all possessed three β -strands and an α -helix (Figure 3.16). Previous analysis of expression values of genes up-regulated from macroschizont to merozoite stage ApiAP2 genes identified four AP2 domain genes with a fold change greater than 2: TA13515, TA16485, TA11145 and TA12015 (see Chapter 2). Orthologues of these domains were identified using BLAST and the NCBI Search Tool for Conserved Domains, and the results are presented in Table 3.5. In addition, putative target DNA core motifs of respective ApiAP2s and motif-stage activity (predictive effect of the AP2 motif on expression of target genes at certain life cycle stage) were obtained from published data for *P. falciparum* (Campbell et al., 2010) and *C. parvum* (Oberstaller et al., 2013). A threshold of a minimum of 50% sequence identity over the ApiAP2 domains was used to identify *T. annulata* ApiAP2 orthologues. TA13515, TA12015, TA16485 and TA11145 proteins and their AP2 domain sequences were then aligned in order to identify shared regions/motifs between orthologues from *T. annulata*, *T. parva* and *T. orientalis*. The AP2 alignment (Figure 3.17) confirmed strong conservation of the domain within the *Theileria* genus. Residues corresponding to the three β -strands and the α -helix of the AP2 domain displayed the highest level of conservation in comparison to the rest of the protein sequence (data not shown). Moreover, the level of conservation of a particular AP2 domain (e.g. in TA11145) was greater across the three species relative to other distinct AP2 domains within a species (e.g. TA11145 vs TA13515) (Figure 3.17).

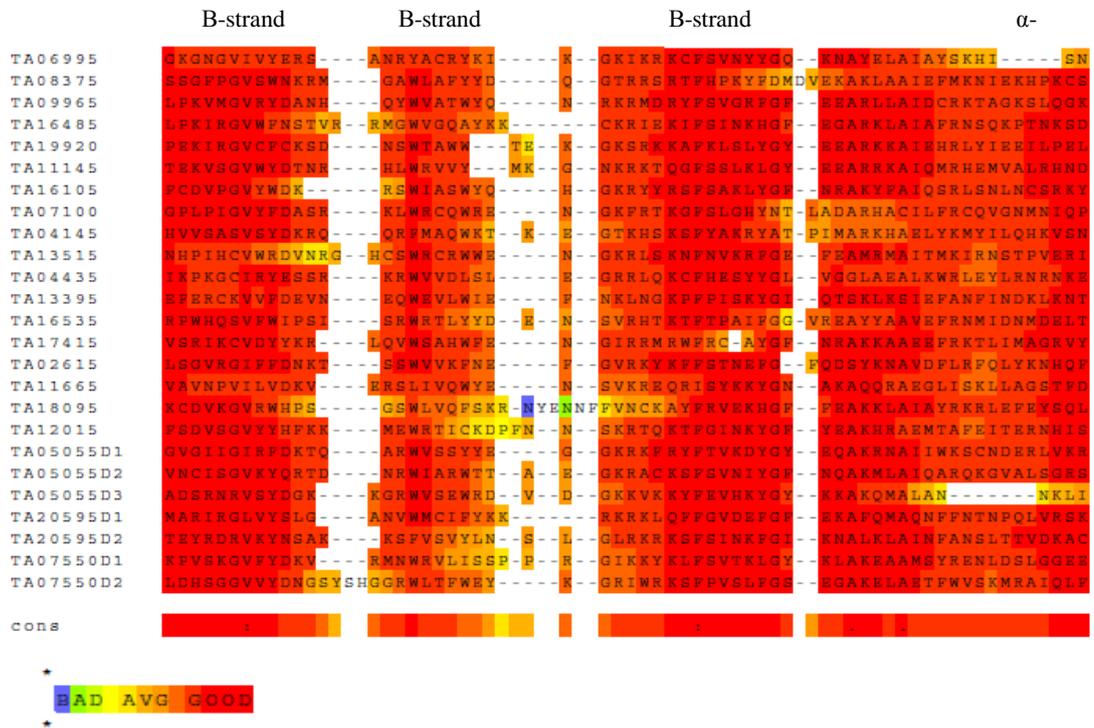


Figure 3.16. Domain alignment of all *T. annulata* ApiAP2 proteins using T-Coffee. Strong conservation of three β -strands and an α -helix is visible (red). D1, D2, D3 - refers to the AP2 domain number in the protein numbered from N-to C-terminus.

<i>T.annulata</i>	<i>T.parva</i>	<i>T.orientalis</i>	<i>P.falciparum</i>	<i>P.berghei</i>	<i>B.bovis</i>	<i>C.parvum</i>	motif activity	core motif
TA13515	TP02_0497	TOT_020000484	PFL1085w	PB000234.02.0	BBOV_II005480	-	gametocytes	GTAC
TA16485	TP01_1126	TOT_010001070	PFL1900w	PB300626.00.0	BBOV_IV011830	-	gametocytes	TCTA
TA11145	TP04_0872	TOT_040000066	MAL8P1.153	PB300504.00.0	BBOV_III009600	-	gametocytes, sporozoite	ACACAC
TA12015	TP02_0226	TOT_020000208	-	-	BBOV_III003770	cgd8_810	gametocytes, zygote	GGGG

Table 3.5. Orthologues of the ApiAP2 gene up-regulated in *T. annulata*. Putative target DNA core motif of respective ApiAP2s are highlighted in pink; motif stage activity is based on data from *P. falciparum* (Campbell et al., 2010) and *C. parvum* (Oberstaller et al., 2013). A minimum of 50% sequence identity between ApiAP2 domains was used as a cut-off for the identification of orthologues of the *T. annulata* ApiAP2 domains.

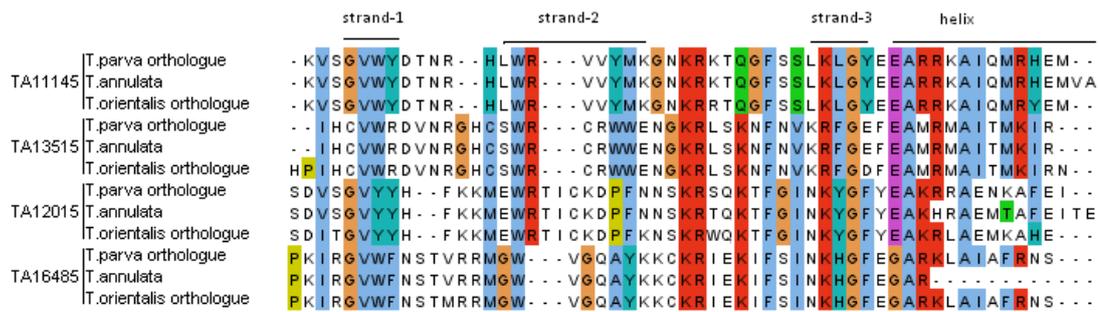


Figure 3.17. Alignment of the ApiAP2 domain of the four up-regulated *T. annulata* ApiAP2 genes (TA11145, TA13515, TA12015, TA16485) with their orthologues in *T. parva* and *T. orientalis*. The domain consists of three β -strands and an α -helix and shows strong conservation across the three *Theileria* species.

Additionally, each of the four selected *Theileria* ApiAP2 domains (TA13515, TA11145 and TA16485) was aligned to its potential *P. falciparum* orthologue. The TA13515 and TA16485 ApiAP2 domains showed stronger conservation to the *Plasmodium* domain (Figure 3.18) in comparison to the others (TA11145 and TA12015 - Figure 3.19. and Figure 3.21.) suggesting similar binding specificities of these domains to their DNA motifs across these genera. The TA11145 ApiAP2 domain also showed a significant degree of conservation with the *P. falciparum* orthologue, particularly over beta strand 1 and 2 and the alpha helix. No *P. falciparum* orthologue of TA12015 was identified either in this study or that of Balaji et al. (2005). However, *T. annulata*, *T. parva* and *T. orientalis* orthologues of the TA12015 ApiAP2 domain were aligned to a predicted *C. parvum* orthologue. Analysis across this domain showed conservation in particular amino acid sequences, suggesting the binding specificities reported for *C. parvum* domain may be similar for *Theileria* species (Figure 3.21).

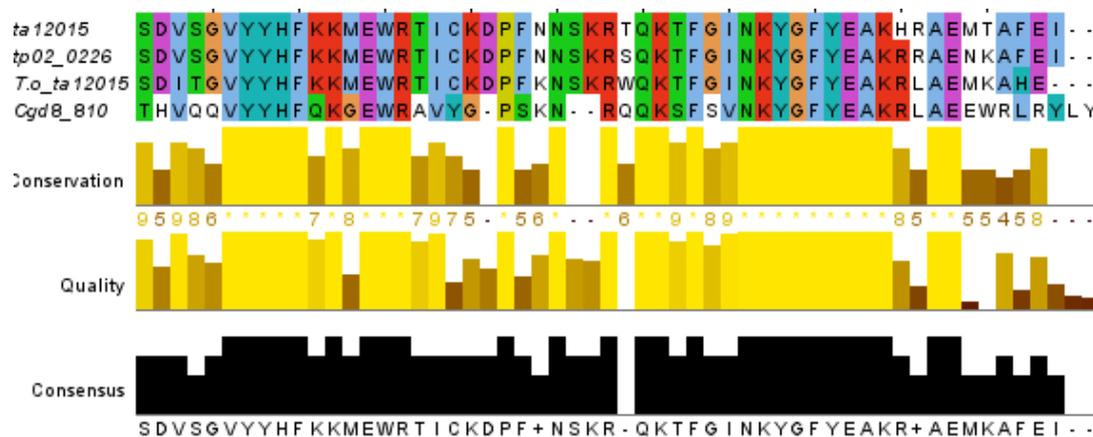


Figure 3.21. ClustalW2 alignment of ApiAP2 domain of TA12015 in *T. annulata* and orthologues in *T. parva*, *T. orientalis* and *C. parvum*. Conservation level of particular amino acid substitution across the species (92 % identity) is visible with moderate conservation across the two genera (53%).

3.3.6 Motif distribution in the genome of *T. annulata* and enrichment analysis of motifs bound by *P. falciparum* ApiAP2s

The distribution of motifs previously identified as targets for *Plasmodium* ApiAP2 proteins (Campbell et al., 2010; Painter et al., 2011) was investigated within putative promoter (IGRs) regions upstream of *T. annulata* genes. Based on the average size of the *T. annulata* intergenic region, a length restriction of ~400 bp was used in a search to capture relevant regulatory motifs. Regions of 400bp, 300bp, 200bp, 100bp upstream of the ATG translation start site and also within the first 200bp of coding sequence of the gene were interrogated using DNA Motif Pattern Search of PiroplasmaDB, with a focus on motifs potentially recognised by *Plasmodium* orthologues of three macro-mero up-regulated TaApiAP2 (TA11145, TA12015, TA16485) factors, and the orthologue of the *Plasmodium* AP2-G factors linked to gametocytogenesis (Sinha et al., 2014; Kafsack et al., 2014) (Table 3.6). All previously identified *Plasmodium* motifs were found in the *T. annulata* genome, but no conserved distribution of the motifs in particular distance categories of upstream sites was observed. The inverse of the ‘CACACA’ motif (GTGTGT) was previously recognised as the most over-represented motif throughout non-coding regions in *T. annulata* and *T. parva* (Guo and Silva, 2008) and a truncated version (GTGTG) was also identified by MEME (see section 3.3.1) in the upstream region of *T. annulata* genes. It can

be concluded that motif distribution is either random or further analysis of specific gene clusters is required to identify potential ApiAP2 target genes. Individual motif enrichment analysis was performed for G-box, 'GTGTAC', 'TCTACA' and 'ACACAC' AP2 motifs using the custom Perl script and assessed on the basis of a two-tailed p-value (Fisher's exact test) to calculate whether there is a significant association between motif occurrence and gene set of interest. Results for individual target DNA motifs of TA13515, TA11145, TA12015 and TA16485 are detailed in section below.

Domain ID in <i>Plasmodium</i>	Motif	no. of total hits in <i>T.annulata</i>	400-300bp upstream on the same strand	300-200bp upstream	200-100bp upstream	100-0bp upstream	0-100bp	100-200bp
PF07_0126_DLD	A(T)A(T)TTTCC	10630	199	253	284	285	209	159
PF14_0633	T(C)G(A)CATGC(T)A(G)	1626	39	37	33	42	24	22
PF13_0267	AT(A)T(C)C(T)TAG(A)A(T)A(T)	21033	657	742	931	1067	416	424
PF14_0079	TG(A)C(A)A(C)ACCA(G)	4761	86	93	82	83	96	121
PF11_0404_D1	T(C)AGAA(T)C(A)AA	5637	171	123	137	123	162	154
PF13_0097	T(C)A(T)G(A)C(A)TCAG(A)A(G)	7190	171	161	130	130	167	192
PF13_0235_D1	C(T)G(A)C(A)GGGG(A)C(A)C(T)	1407	30	22	19	17	36	35
PF10_0075_D1	GG(T)GTCGACC(A)C	12	0	0	2	0	0	0
PF10_0075_D2	T(C)C(A)TTGCC	1861	28	23	21	27	41	33
PF10_0075_D3	GTGCAC(T)T(A)A	436	9	9	13	10	11	8
PFL1075w	T(A)A(T)TATAT(A)A(T)	49720	1188	1394	1669	1866	721	773
PFL1900w_DLD	TCTA(C/T)A(A/G)A	2470	68	60	57	76	51	50
PFE0840c_D2	(T/G)GA(C/T)ATC(A/T)	2830	53	36	35	24	57	48
PF14_0533	CA(G)CACACAC(A)	124	2	3	4	4	1	2
PFF0670w_D1	A(T)T(C)A(G)AG(A)C(A)C(A)C(T)A(G)	33214	771	730	757	609	705	728
PFF0670w_D2	C(A)T(C)C(A)TAG(T)A(G)G(T)	19646	423	429	500	508	306	346
PFF200c_DLD	GGTGCACC(T)	75	2	3	0	0	0	4
MAL8P1.153	C(T)A(G)CACAC(T)A(T)	5966	184	245	288	265	130	174
PF11_0091	AGC(A)ATAC	2108	56	60	30	21	0	0
PF11_0442	AGCTAGCT	70	1	2	3	7	0	1
PF0985w_D1	C(T)A(G)CAC(T)AC(T)	15653	404	457	505	485	316	372
PF0985w_D2	GTGTTACAC	49	2	0	3	6	2	3
PF13_0026	T(C)G(A)CACACA(G)C(A)	483	9	10	11	6	6	10
PFL1085w	G(A)TGT(G)A(C)CAC(T)	2870	48	40	35	56	67	55

Table 3.6. *P. falciparum* ApiAP2 target gene motifs and distribution in *T. annulata* upstream regions (blue denotes - three up-regulated *T. annulata* ApiAP2 orthologues - PFL1085w - TA13515, MAL8P1.153 - TA11145 and PFL1900w_DLD - TA16485).

3.3.6.1. G-box/C-box motif analysis

Enrichment analysis of the G-box motif, which is recognised by the *C. parvum* orthologue of TaApiAP2-TA12015, showed enrichment within 400bp upstream of the TSS on the positive strand with 45.8% of merozoite to piroplasm up-regulated genes (11 out of 24 genes) possessing the motif and 33.33% (8 out of 24 IGRs) for the negative strand (Table 4.8). Within 100bp upstream of the TSS, 20.8% of merozoite to piroplasm up-regulated genes were found to possess this motif, suggesting a shared position in the majority of this subset of up-regulated genes and an association with the piroplasm stage of the life-cycle. In contrast, the motif was significantly under-represented upstream of genes down-regulated from macroschizont to merozoite, present only in 1.77% of IGRs on the positive strand. Because this G-box motif is an inverse of a 4C-box, it was also found on the negative strand, within 400bp upstream of the TSS of 82.35% of TashAT genes. As expected, no statistically significant enrichment was found for SVSP genes since the longer 5C-box motif was previously identified by MEME to be associated with this down-regulated gene family. These results were statistically significant, p-value ≤ 0.01 .

As a G-box motif (inverse of a C-box) was previously found to be associated with Heat Shock Protein (HSP) genes in *P. falciparum* (Militello et al., 2004) and *C. parvum* (Cohn et al., 2010), a MEME analysis was performed for *T. annulata* HSP genes. The presence of both G-box and C-box was confirmed in the upstream regions of these genes. A C-box was found in intergenic regions on the positive strand and as an inverse of G-box on the negative strand of eight HSPs (Figure 3.22): TA07065, TA14920, TA06845, TA10500, TA12100, TA12105, TA10720, TA11610. Additionally, the PiroplasmaDB DNA Motif Pattern Search tool identified the motif, (A/G)NGGGG(C/A) in 8 out of 34 *T. annulata* HSP genes, 400bp upstream of the TSS on the positive strand and in 5 on the negative strand (Table 3.7). The enrichment was shown to be significant, p-value ≤ 0.01 .

Name	Strand	Start	p-value	Sites [?]
TA07065	+	349	8.40e-06	TGATGTTGAT CCCCAC TAATAGAGTT
TA12105	-	209	8.40e-06	AAGTGTCAAC CCCCAC TGATGCTATT
TA12100	+	917	8.40e-06	AAGTGTCAAC CCCCAC TGATGCTATT
TA10720	+	266	8.40e-06	ATATTCCTGT CCCCAC CATAAAACCA
TA14920	+	521	8.40e-06	CCCAATGGCT CCCCAC AAGCCAGGAA
TA10500	-	327	3.59e-05	ATCACTATAT CCCCAT CAATGGAAAA
TA06845	-	191	3.59e-05	ATAGTTCGAT CCCCAT GGCTACTTGA
TA11610	+	471	4.43e-05	TCAACTGATA CCCCAG TTATGAGGAC
TA06470	-	68	9.93e-05	AGTATTTAAA CCACAC ATTTATTTTT
TA20640	+	77	1.89e-04	AAAAACAATT TCCCAT GTAATATAAT
TA09395	-	55	1.89e-04	TTCTTCTTAT TCCCAT AAACCTCCAT

Figure 3.22. The C-box motif identified in the *T. annulata* Heat Shock gene family by MEME analysis (C-box - highlighted in blue).

Table 3.7. Statistical enrichment analysis of (A/G)NGGGG(C/A) motif within *T. annulata* intergenic regions (400 bp upstream of the translation start site). Statistically significant data (P value < 0.05, Fisher's Exact Test) is highlighted in yellow.

[A/G]NGGGG(C/A) motif	TashAT genes	non-TashAT genes	macro-mero most down-regulated genes	non-macro-mero most down-regulated genes	macro-mero most up-regulated genes	non-macro-mero most up-regulated genes	mero-piro most up-regulated genes	non-mero-piro most up-regulated genes	mero-piro most down-regulated genes	non-mero-piro most down-regulated genes	SVSP genes	non-SVSP genes	HSP genes	non-HSP genes
400 bp upstream positive strand	0% (0/17) P=0.3950	9.3% (349/3751)	1.77% (2/113) P=0.0025	9.49% (347/3655)	6.75% (10/148) P=0.3836	9.36% (339/3620)	45.8% (11/24) P<0.0001	9.03% (338/3743)	5% (1/20) P=1	9.28% (348/3747)	0% (0/48) P=0.0200	9.38% (349/3719)	23.53% (8/34) P=0.0106	9.13% (341/3733)
400 bp upstream negative strand	82.35% (14/17) P<0.0001	10.66% (400/3751)	14.15% (16/113) P=0.2834	10.9% (398/3655)	15.54% (23/148) P=0.0802	10.8% (391/3620)	33.33% (8/24) P=0.0030	10.8% (406/3743)	15% (3/20) P=0.4766	10.9% (411/3747)	8.3% (4/48) P=0.8151	11.02% (410/3719)	14.7% (5/34) P=0.2614	10.9% (409/3733)
200-100bp upstream positive strand	0% (0/17) P=1	2.42% (91/3751)	0.89% (1/113) P=0.5256	2.46% (90/3655)	2.7% (4/148) P=0.7815	2.4% (87/3620)	37.5% (9/24) P=0.0001	2.2% (82/3743)	0% (0/20) P=1	2.42% (91/3747)	0% (0/48) P=0.6306	2.44% (91/3719)	11.76% (4/34) P=0.0085	2.33% (87/3733)
100-0bp upstream positive strand	0% (0/17) P=1	2.1% (79/3751)	0% (0/113) P=0.1748	2.16% (79/3655)	0.67% (1/148) P=0.3728	2.15% (78/3620)	0% (0/24) P=1	2.11% (79/3743)	0% (0/20) P=1	2.1% (79/3747)	0% (0/48) P=0.06259	2.12% (79/3719)	0% (0/34) P=1	2.11% (79/3733)

3.3.6.2. GTGTAC, TCTACA and ACACAC motif analysis

Motif enrichment analysis of the 'GTGTAC' motif recognised by the *Plasmodium* AP2-G orthologue of TaApiAP2, TA13515, showed that within 400bp upstream on the positive strand, the most highly up-regulated merozoite to piroplasm genes were enriched for this motif (29.16%). A P-value = 0.0001 was considered as highly statistically significant. No statistical enrichment of this motif was found in any other stage regulated genes or in the upstream regions of TashAT, SVSP or HSP genes, implying this motif is of importance for genes up-regulated from merozoite to piroplasm: the precursor to/or the gametocyte stage of the *Theileria* life-cycle (Melhorn and Schein, 1984).

Motif enrichment analysis of the TCTA[C/T]A[A/G]A motif bound by the *Plasmodium* orthologue of TaApiAP2-TA16485, revealed the presence of this motif within 400bp upstream on the positive strand of 41.2% (7 out of 17) TashAT family genes, and 51.9% (9 out of 17) of TashAT genes on the negative strand. Both results are highly statistically significant, p-value \leq 0.001. The 'TCTACA' motif was also found to be enriched for macroschizont to merozoite down-regulated genes (11.5%), in general, 400bp upstream of the TSS on the positive strand, p-value = 0.0574 (Table 3.9). No statistical enrichment of this motif was found in the upstream regions of SVSP family and HSP genes.

Motif enrichment analysis of the 'CACACA' motif bound by the *Plasmodium* orthologue of TaAPiAP2-TA11145 showed that within 400bp upstream on the positive strand, the motif was under-represented in IGRs of macroschizont to merozoite down-regulated genes. Only 13.27% (15 out of 113) of these genes possess the motif and this corresponds with a statistically significant under-representation (p-value \leq 0.01). In contrast, evidence of enrichment was found in IGRs of macroschizont to merozoite up-regulated genes (31.1%, 46 out of 148; p-value = 0.0631) and merozoite to piroplasm up-regulated genes (45.8%, 11 out of 24; p-value = 0.028; Table 3.10). No significant enrichment or under-representation was evident in IGRs representing the TashAT or SVSP gene families.

GTGTAC motif	TashAT genes	non-TashAT genes	macro-mero most down-regulated genes	non-macro-mero most down-regulated genes	macro-mero most up-regulated genes	non-macro-mero most up-regulated genes	mero-piro most up-regulated genes	non-mero-piro most up-regulated genes	mero-piro most down-regulated genes	non-mero-piro most down-regulated genes	SVSP genes	non-SVSP genes	HSP genes	non-HSP genes
400 bp upstream positive strand	0% (0/17) P=1	4.66% (175/3751)	3.54% (4/113) P=0.81	4.67% (171/3655)	6.08% (9/148) P=0.42	4.58% (166/3620)	29.16% (7/24) P=0.0001	4.48% (168/3743)	0% (0/20) P=1	4.67% (175/3747)	2.08% (1/48) P=0.72	4.67% (174/3719)	0% (6/34) P=0.40	4.68% (175/3733)
400 bp upstream negative strand	11.76% (2/17) P=0.19	4.71% (177/3751)	3.54% (4/113) P=0.81	4.78% (175/3655)	5.40% (8/148) P=0.69	4.72% (171/3620)	0% (0/24) P=0.62	4.78% (179/3743)	0% (0/20) P=0.62	4.77% (179/3747)	4.16% (2/48) P=1	4.75% (177/3719)	5.88% (2/34) P=0.40	4.74% (177/3733)

Table 3.8. Statistical enrichment analysis of GTGTAC motif within *T. annulata* intergenic regions (up to 400 bp upstream of the translation start site). Statistically significant data (P-value < 0.05, Fisher's Exact Test) is highlighted in yellow.

TCTA[C/T]A[A/G]A motif	TashAT genes	non-TashAT genes	macro-mero most down-regulated genes	non-macro-mero most down-regulated genes	macro-mero up-most regulated genes	non-macro-mero most up-regulated genes	mero-piro up-most regulated genes	non-mero-piro most up-regulated genes	mero-piro most down-regulated genes	non-mero-piro most down-regulated genes	SVSP genes	non-SVSP genes	HSP genes	non-HSP genes
400 bp upstream positive strand	41.17% (7/17) P<0.0001	6.71% (252/3751)	11.5% (13/113) P=0.0574	6.7% (246/3655)	4.05% (6/148) P=0.1873	6.98% (253/3620)	4.16% (1/24) P=1	6.8% (258/3743)	10% (2/20) P=0.6441	6.85% (257/3747)	2.08% (1/48) P=0.2554	6.93% (258/3719)	5.88% (2/34) P=1	6.8% (257/3733)
400 bp upstream negative strand	52.94% (9/17) P<0.0001	7.14% (268/3751)	9.73% (11/113) P=0.2742	7.3% (266/3655)	6.75% (10/148) P=1	7.4% (267/3620)	4.16% (1/24) P=1	7.37% (276/3743)	15% (3/20) P=0.1781	7.3% (274/3747)	10.41% (5/48) P=0.2587	7.3% (272/3719)	8.82% (3/34) P=0.7358	7.34% (274/3733)

Table 3.9. Statistical enrichment analysis of TCTA[C/T]A[A/G]A motif within *T. annulata* intergenic regions (up to 400 bp upstream of the translation start site). Statistically significant data (P-value < 0.05, Fisher's Exact Test) is highlighted in yellow.

[C/T][A/G]CACA[C/T][A/T] motif	TashAT genes	non-TashAT genes	macro-mero most down-regulated genes	non-macro-mero most down-regulated genes	macro-mero most up-regulated genes	non-macro-mero most up-regulated genes	mero-piro most up-regulated genes	non mero-piro most up-regulated genes	mero-piro most down-regulated genes	non-mero-piro most down-regulated genes	SVSP genes	non-SVSP genes
400 bp upstream positive strand	11.76% (2/17) P=0.3934	24.47% (918/3751)	13.27% (15/113) P=0.0038	24.76% (905/3655)	31.1% (46/148) P=0.0631	24.1% (874/3620)	45.8% (11/24) P=0.0280	24.3% (909/3743)	40% (8/20) P=0.1177	24.33% (912/3747)	16.66% (8/48) P=0.2396	24.52% (912/3719)
400 bp upstream negative strand	11.76% (2/17) P=0.2721	24.58% (922/3751)	21.23% (24/113) P=0.4394	24.62% (900/3655)	27% (40/148) P=0.4950	24.4% (884/3652)	16.6% (4/24) P=0.4792	24.57% (920/3743)	15% (3/20) P=0.4382	24.58% (921/3737)	14.58% (7/48) P=0.1284	24.65% (917/3719)

Table 3.10. Statistical enrichment analysis of [C/T][A/G]CACA[C/T][A/T] motif within *T. annulata* intergenic regions (up to 400 bp upstream of the translation start site). Statistically significant data (P-value < 0.05, Fisher's Exact Test) is highlighted in yellow.

3.3.7 Motifs found in upstream regions of ApiAP2 genes of *Theileria*

In *Plasmodium*, ApiAP2 genes form an interaction network with themselves (Campbell et al., 2010), which suggests that they may also regulate their own expression in *Theileria* species. Analysis of the upstream regions of the four selected up-regulated (macroschizont to merozoite) ApiAP2s revealed the presence of their predicted target DNA motifs in their own IGR (Table 3.11) suggesting auto-regulation of these putative transcription factors, in addition to putative regulation of transcription of other target genes including other ApiAP2 genes. A complete alignment of upstream sequences of *T. annulata* and *T. parva* ApiAP2s can be found in the Appendix (3.8-3.11). Minimum four-nucleotide core sequences with variable flanking nucleotides were taken into account (based on *Plasmodium* binding specificities; Campbell et al., 2010) in searching for motifs upstream of *T. annulata* ApiAP2 genes. Most notable were the upstream region of TA13515, with three copies of the core 'GTAC' motif identified, and TA11145 where seven copies of CACACA/ACACAC motif were found: TA12015 IGR has one copy of the C-box (CCCC motif) and TA16485 two copies of the TCTA[C/T] motif the ApiAP2 domain is predicted to bind.

ApiAP2/Motif	<u>[A/G]TGTACA</u> [C/T]	[C/T] <u>ACACAC</u> [A/T]	G-box/C-box	<u>TCTA</u> [C/T]AA
TA13515	3	1	4	1
TA11145	1	7	2	-
TA12015	1	2	1	-
TA16485	1	3	-	2

Table 3.11. Number of DNA motifs (with a minimum of 4 core nucleotides) found in the upstream regions of the four macroschizont to merozoite up-regulated ApiAP2 genes. Underlining represents the core nucleotides involved in specific protein-DNA binding interaction defined for orthologous ApiAP2 domains in *P. falciparum*. The number of motifs in the upstream region of the gene that encodes the ApiAP2 motif-binding domain is highlighted in yellow.

3.3.8 Temporal expression pattern analysis of ApiAP2 genes and their potential target genes

Comparison of expression profiles of four ApiAP2 genes (TA11145, TA12015, TA16485 and TA13515) with profiles of predicted target genes (plotted as a mean value for all genes identified) containing the motif bound by the AP2 domain was performed.

The Pearson correlation coefficient value measures the strength of the relationship between two variables. A highly significant negative correlation (≤ -0.9) was observed for TA11145 and genes most down-regulated from macroschizont to merozoite gene with a 'CACACA' motif in their IGR (Figure 3.23.A). A similar correlation was observed for the ApiAP2 encoding TA16485 gene and TashAT genes possessing the TCTA[C/T]A motif (Figure 3.23.B) in their IGR and the most down-regulated macroschizont to merozoite dataset showing enrichment for this motif in the IGRs (Figure 3.23.D). In a similar manner, a negative correlation between expression of TA12015 and TashAT genes possessing the C-box motif (Figure 3.23.C) was found. Perfect positive correlation (=1) was identified for TA13515 and merozoite to piroplasm up-regulated genes (Figure 3.24.C). Moreover, a highly significant positive Pearson

correlation (≥ 0.9) was observed between TA11145 and the macro-mero most up-regulated gene set enriched for 'CACACA' in their IGR (Figure 3.24.A), demonstrating a strong relationship between these putative target gene sets and the ApiAP2 factors, this correlation suggesting the ApiAP2 factors operate to activate these set of genes. A strong negative correlation was also found for expression patterns of TA12015 and HSP family genes with a G-box motif (Figure 3.23.E). In contrast, the mean expression profile of mero-piro up-regulated genes with a G-box showed a low correlation with the expression pattern of the gene TA12015 encoding the domain (Figure 3.24.B) predicted to bind to this motif. A similar lack of correlation was indicated for the mero-piro up-regulated genes with the 'CACACA' motif (Figure 3.24.D) predicted and the ApiAP2 TA11145. This may suggest that other regulatory factors are involved regulating expression of these genes.

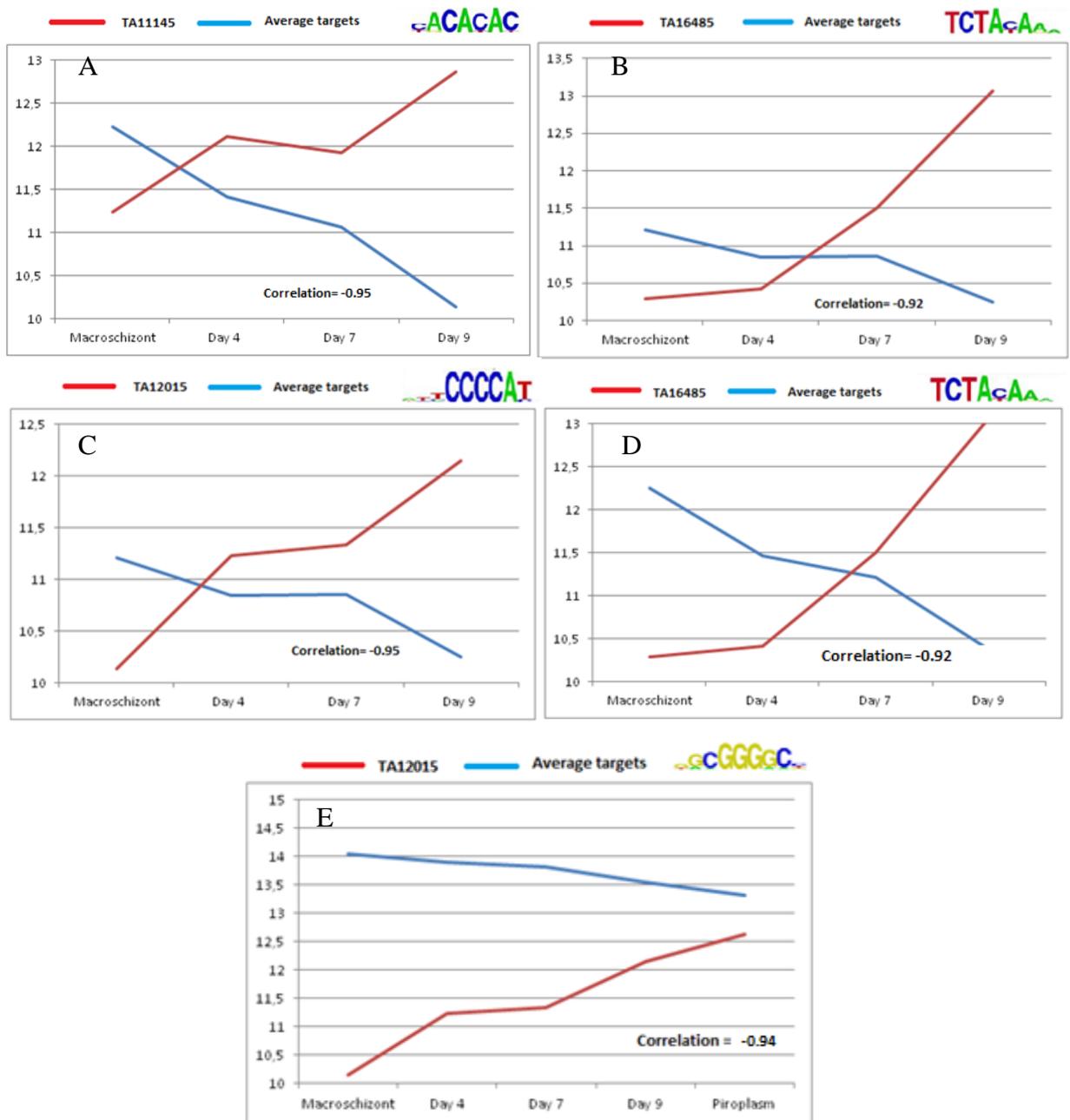


Figure 3.23. Negative correlation plots of ApiAP2 gene mRNA abundance (in red) and putative target genes possessing the motif bound by the AP2 domain (in blue). Target plot line represents the average mRNA abundance profiles at each time point based on *T. annulata* microarray data. The Pearson correlation values are presented at the bottom right of each plot box. A - TA11145 and macro-mero most down-regulated genes with a CACACA motif; B - TA16485 and TashAT genes with a TCTAC motif; C - TA12015 and TashAT genes with a C-box motif; D - TA16485 and macro-mero most down-regulated genes with a TCTAC motif; E - TA12015 and Hsp genes with a G-box motif.

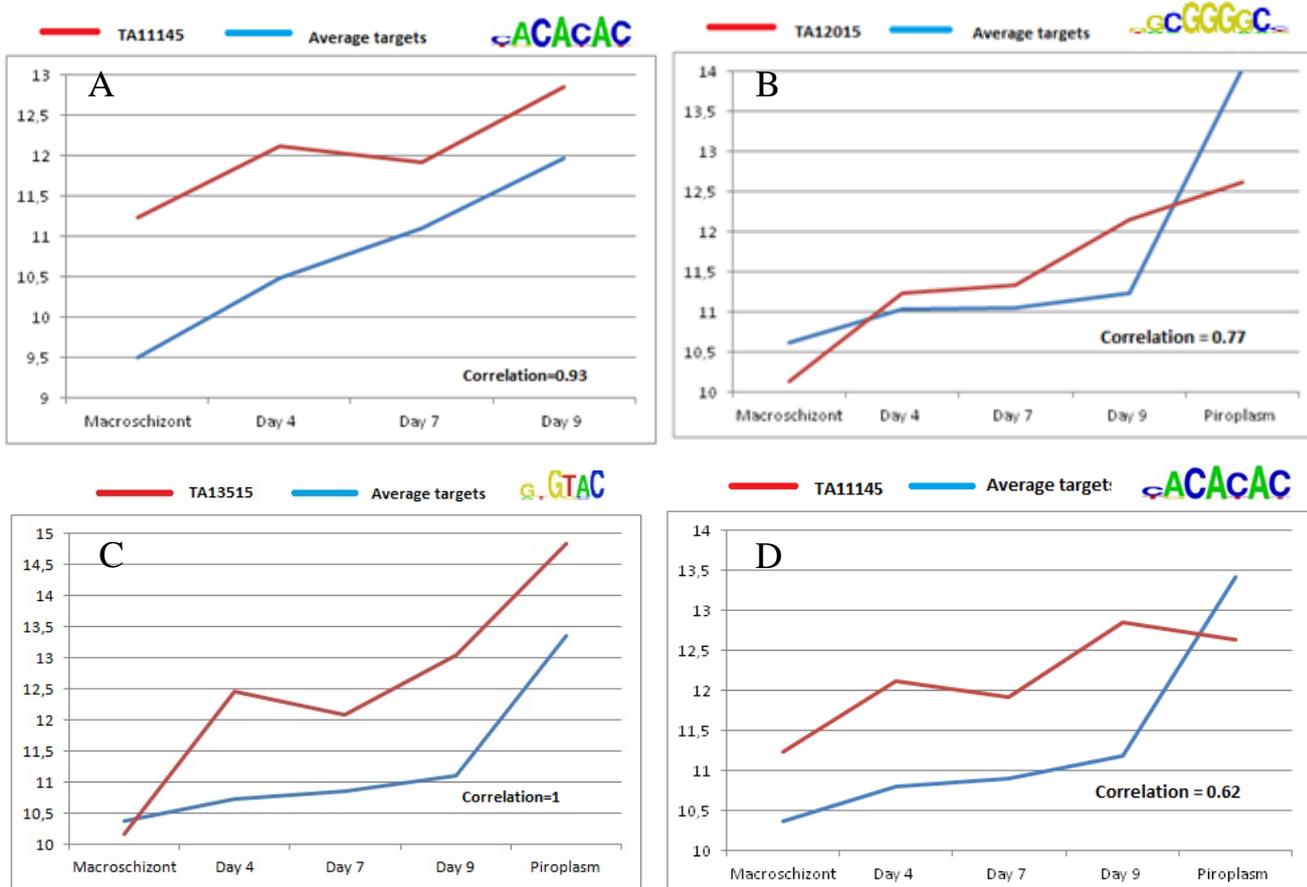


Figure 3.24. Positive correlation plots of ApiAP2 gene mRNA abundance (in red) and putative target genes possessing the motif bound by the AP2 domain (in blue). Average targets represent the average mRNA abundance profiles at each time point based on *T. annulata* microarray data. The Pearson correlation values are presented at the bottom right of each plot. A - TA11145 and macro-mero most up-regulated genes with a CACACA motif; B - TA12015 and mero-piro most up-regulated genes with a G-box motif; C - TA13515 and mero-piro up-regulated genes with GTGTAC motif ; D - TA11145 and mero-piro up-regulated genes with a CACACA motif.

3.4. Discussion

Bioinformatic approaches for *de novo* identification of transcription binding sites and other gene regulatory elements have been proved to be successful for creating high quality hypotheses of transcriptional control mechanisms (Guo and Silva, 2008; Mullapudi et al., 2009; Oberstaller et al., 2013). Since the sequence information from closely related species can be a starting point for the identification of evolutionarily conserved and biologically important regions of DNA (Zhang and Gerstein, 2003), the investigation of the presence of putative orthologues ApiAP2 TFs and their target motifs, previously identified for *P. falciparum* (Campbell et al., 2010) and *C. parvum* (Oberstaller et al., 2013), was performed for *Theileria*. The work presented in this chapter identified *de novo* conserved and over-represented motifs in the upstream regions of functionally related and/or co-expressed genes of *T. annulata*.

MEME screening of *T. annulata* sets of intergenic regions corresponding to stage-specifically regulated genes and gene families revealed only a small number of shared nucleotide motifs. Since gene families or clusters of similarly expressed genes do not always share the same cis-elements (Kundaje et al., 2007), the prediction of DNA motifs can be a difficult process. Discovery of only a small number of statistically significant motifs may also be due to small differences in the flanking regions of a DNA motif that prevent the sequence from being considered as significant. A limited number of *de novo* identified motifs were considered as statistically significant in this study and interesting from the biological point of view, i.e. 4C- and 5C-rich motifs, a G-rich motif, and CACACA/TGTGTG motifs. Forms of these motifs have previously been identified in the genome of *Theileria* (Guo and Silva, 2008) and other Apicomplexan species (Campbell et al., 2010, Oberstaller et al., 2013).

It is worthwhile noting that Guo and Silva (2008) used a different strategy for motif discovery, by searching for motifs within 300 bases of IGRs of all *T. parva* and *T. annulata* genes, while I focused on upstream regions of functionally related genes and genes co-expressed from the macroschizont to the merozoite stage. All the motifs detailed in my study showed a likelihood of functionality based on conservation with other *Theileria* species and also similarity with

motifs identified in *P. falciparum*. Precise localisation of the DNA binding site in relation to the translation start site was also considered in the present analysis, as functional motifs are usually found between around 100 and 150 bp upstream of the TSS (Hughes et al., 2000; Iengar and Joshi, 2009). While intergenic upstream sequences used in this study had a variety of lengths, a strong bias of positional location relatively close to the ATG start codon was often observed, strongly suggesting functional involvement in expression of genes enriched for these motifs in their upstream region. In addition, evidence for significant enrichment (p -value <0.01) in sets of genes with similar expression patterns indicate these motifs have a potential role in determining differential gene expression.

A C-box motif was identified as enriched in 5' intergenic regions - almost unique - of *T. annulata* genes that are down-regulated from the macroschizont to the merozoite stage. The majority of the most down-regulated genes highlighted by this analysis map to gene families predicted to encode proteins that constitute the macroschizont secretome (SVSPs and TashAT family) and thus encode a signal peptide. This supported Guo and Silva (2008) who showed that a C-box motif is associated with genes involved in protein synthesis and genes with a signal peptide. The motif is similar to the consensus binding site for C2H2 zinc finger protein involved in cellular proliferation and oncogenesis (Guo and Silva, 2008; Morris et al., 1994). Moreover, in *Plasmodium* C-rich motifs were identified in upstream regions of mitochondrial genes (Iengar and Joshi, 2009) and a 'CCCCTTA' motif was found enriched 700 to 999 bases upstream of the start codon of the genes highly expressed during phases of rapid cellular proliferation (Young et al., 2008). This provides an interesting parallel to the association of the C-box motif with genes predicted to be involved in the transformed phenotype of the *Theileria* infected cell, upon which the parasite depends for cellular division and tissue dissemination (Shiels et al., 2006).

TashAT family genes function as modulators of host cell phenotype (Shiels et al., 2004). Phylogenetic and gene synteny analysis of TashAT genes in *T. annulata* and the TphN family in *T. parva* indicated that their common ancestor had a condensed gene family that expanded and diversified following the *T. annulata*/*T. parva* split (Shiels et al., 2006). MEME analysis of the cluster across

these two species identified an enlarged 4C-box motif (TCCCCAT) that is present in the majority of the IGRs of TashAT genes in both species. The TashAT family has probably evolved by tandem gene duplication, and this could account for identification of the motif also in the 3' IGR analysis, i.e. it is actually upstream of the next member of the cluster. Furthermore, the fact that protein-coding regions have diverged considerably within the TashAT cluster and between orthologues of *T. annulata* and *T. parva* suggests that the strong positional conservation of this motif upstream of the majority of TashAT and TphN genes is of functional significance.

To further determine if the 4C-box motif is associated with down-regulation of genes in the cluster or with the TashAT family in particular, analysis of the intergenic region of the gene encoding the Tash-a (TA03110) protein was performed. Tash-a is the only gene in the cluster that is conserved in the non-transforming *Theileria* species, *T. orientalis*, and is up-regulated during merogony, based on microarray data and IFAT (Hyashida et al., 2013). While the larger motif identified for the rest of the TashAT cluster by MEME was absent in the intergenic region of this gene, core C-boxes and G-boxes were identified. Thus, whether a core box of TCCCCAT actually confers a signal for down-regulation is at present unclear and can be considered to be unlikely.

Interestingly, while TashAT genes were associated with the 4C-box motif, the SVSP gene family was shown to be significantly associated with a related 5C-box motif. Like TashATs, predicted SVSP proteins possess signal peptides and are secreted into the host cell compartment (Schmuckli-Mauerer et al., 2009). However, genome localisation of both sets of genes is different and they are predicted to have distinct functions (Weir et al., 2010). Both TashAT and SVSP genes are down-regulated from the macroschizont to the merozoite stage and have similar expression profiles. The presence of similar C-box core-motifs does not guarantee regulation by the same transcription factor(s); however it does imply that it may operate in a mechanism for coordinated down-regulation. Thus the C-box motif can be postulated as a possible binding site whose functionality needs to be determined experimentally. This could be a DNA binding site for specific transcription factors or represents other signals such as accessibility sites for proteins involved in epigenetic regulation. *Theileria orientalis*, a non-

transforming *Theileria*, lacks the SVSP family and possesses only one TashAT/TpHN-like gene (Hayashida et al., 2013). The IGR of this gene contains a core 4C-box but with significantly different flanking nucleotides. This could suggest that a more primitive C-box motif has been co-opted to regulate the expression of genes involved in manipulating the host cell phenotype and that it has been retained as the secretome gene families expanded in *T. annulata* and *T. parva*. Further investigation of C-box motifs and the genes they are associated with in non-transforming *Theileria* species may be of interest. Inhibiting expression (with the use of RNAi) of genes required for host cell proliferation could be a viable target for development of novel therapeutics against tropical theileriosis and ECF.

A second motif was identified in a smaller number of genes across the TashAT/TpHN family, a small G-box motif (TAGGGTA) that encompasses the 'TAGGG' gel-shift motif of the up-regulated Tams1 merozoite gene (Shiels et al., 1999). Moreover, a 4G-rich motif was found in upstream regions of *T. annulata* merozoite to piroplasm most up-regulated genes on both positive and negative strands. This motif previously has been associated with ribonucleotide synthesis (Young et al., 2008) and its presence was also confirmed in upstream regions of *Theileria* HSP genes (Millitello et al., 2004). Identification of a similar G-box motif as a binding site of ApiAP2 TFs (Campbell et al., 2010; Oberstaller et al., 2013) suggests the potential importance of this DNA motif in regulating gene expression in *Theileria* parasites.

It is possible that a G-box motif may be recognised as the inverse of the C-box motif associated with down-regulated macroschizont genes, e.g. the TashAT gene family. This presents a potential difficulty as the G-box is more enriched in IGRs of genes that are up-regulated from the macroschizont through to piroplasm. One possibility is that the factors that recognise these motifs have dual function and act as repressors and/or activators of gene expression or there is strand-specificity involved in motif recognition and/or factor function (Young et al., 2008).

Heat shock as a common stress trigger has been linked to initiation of protozoan parasite differentiation events and it may be relevant that a G-box is enriched in

the IGR of HSP genes. However, the expression of heat shock genes from the macroschizont to merozoite stage is constitutive according to microarray data, and data implicating heat shock genes in directly regulating commitment to merogony is lacking for *T. annulata*. Thus, the C/G-boxes may be associated with a general mechanism that allows additional factors to bind, even though it appears to be associated with genes expressed in a differential manner. A G-box motif may also impart properties of intrinsic DNA structure (a special sequence repeated in phase with the DNA helical repeat) involved in organisation of local chromatin structure to stimulate interactions between DNA-binding motifs and transcription factors (Ohyama, 2001). Intrinsic DNA structure may also play a significant role in promoting protein-protein contacts between two transcription factors. The binding process in this case would require distortion of the DNA helix structure either by DNA looping or bending (Kim et al., 1995). Intrinsic DNA bending has been described for many prokaryotic genes operating in both repression and as well activation (Perez-Martin et al., 1994; Perez-Martin and de Lorenzo, 1997).

Since only a small number of motifs were identified *de novo* in *T. annulata*, further analysis focused on searching for similarities between previously described binding motifs sites for AP2 domains of other species that were orthologues of the four genes encoding ApiAP2 domains (TA13515, TA11145, TA12015 and TA16485) significantly up-regulated from the macroschizont to the merozoite/piroplasm stage in *T. annulata*. These genes vary in size and genome localisation, however they all possess a single ApiAP2 DNA-binding domain of around 60 amino acid residues in size. Fortunately orthologues of all four putative DNA binding domains are well described in the literature (Campbell et al., 2010, Oberstaller et al., 2013, Sinha et al., 2014), each with an identified DNA binding site motif (Campbell et al., 2010; Oberstaller et al. 2013). None of the four selected ApiAP2 factors considered in this analysis were completely conserved across all Apicomplexan species. According to the analysis of Balaji et al. (2008): the TA16485 ApiAP2 domain does not possess a direct *C. parvum* orthologue; while for TA13515 there is an orthologue in both *Plasmodium* and *Cryptosporidium*; for TA11145 there are two *C. parvum* ApiAP2 domains that show a phylogenetic relationship with the *T. annulata/Plasmodium* cluster; lastly, no orthologue of the TA12015 ApiAP2 domain was found among

Plasmodium species, but an orthologue was found in *C. parvum*. Differences between the Balaji et al. (2008) phylogeny and the orthologues described in this chapter are likely due to sequence identity thresholds used in the different analyses. Identified ApiAP2 domain orthologues showed stronger identity across genera in comparison to the relationship between the four up-regulated paralogous ApiAP2 domains within *Theileria* species. This provides evidence for recognition of a similar DNA motif by orthologous ApiAP2 domains while different paralogous domains recognise distinct binding sites.

All motifs previously identified for *P. falciparum* ApiAP2 domains (Campbell et al., 2010; Painter et al., 2011) were found to be present in regions of the *T. annulata* genome upstream of coding sequences. However, distribution analysis did not show any specific enrichment within specific regions of upstream sequences. It can be concluded that motif distribution is either random or further analysis of smaller gene sets is required to identify specific local enrichment. Moreover, no significantly enriched, common motifs were found by MEME in the upstream regions of ApiAP2 gene family or other putative transcription factors. This suggests the lack of an essential motif required for gene expression and unique patterns of motifs associated with regulation of different genes.

A statistical enrichment of the *Plasmodium* 'GTGTAC' motif in IGRs of the *Theileria* merozoite to piroplasm up-regulated gene set and a perfect positive expression correlation with putative target genes was of particular interest. The *Plasmodium* ApiAP2 factor (AP2-G) binding to this motif is involved in regulation of sexual development of the parasite (Sinha et al., 2014). Based on the *Theileria* expression pattern of the orthologous ApiAP2 TA13515 gene (see chapter 2) together with the pattern of expression of putative target genes, it is likely that these orthologues regulate gene expression at related points of the life-cycle of the two genera. This might suggest that a commitment step to the precursor of sexual development and sex determination has been conserved across related genera.

Analysis for enrichment of the TCTA[C/T]A motif showed a statistically significant result for upstream regions of TashAT cluster genes. However, it was

not present in upstream region of TA03110 gene. This motif was found adjacent the C-box motif. It is possible that TashAT genes require two or more separate transcription factors for regulation of their expression, which could act as activator and/or repressor or operate as a hetero-dimer involving recognition of two distinct motifs. As the expression pattern of TA16485 and TashAT genes was highly negatively correlated (Pearson correlation = -0.92) it is most likely that this factor acts as a repressor of some of these genes. Statistical enrichment of the 'TCTA[C/T]A' motif was also shown for the most down-regulated macroschizont to merozoite set of genes, and their expression patterns were strongly negatively correlated to TA16485 (Pearson correlation = -0.92). The TA16485 gene was found to have in its upstream region two copies of the 'TCTA[C/T]A' motif the AP2 domain is predicted to bind, one copy of 'GTGTAC' and three copies of the 'CACAC' motif. Thus auto-regulation as well as regulation by other ApiAP2 domain factors is a possibility.

The *C. parvum* orthologue Cgd8_810 (Oberstaller et al., 2013) of the up-regulated, macroschizont to merozoite, TA12015 ApiAP2 gene has been shown to bind to a G-box motif. Based on lack of a *P. falciparum* orthologue, this DNA motif was considered to be a reasonable choice for further investigation of binding specificity of this AP2 domain. The expression pattern of TA12015 was strongly negatively correlated with genes of the TashAT cluster but also showed positive correlation with merozoite to piroplasm up-regulated genes and heat shock genes with a G-box motif. Additionally, an analysis of the upstream region of TA12015 revealed presence of a G-box, 'GTGTAC' and two 'CACAC' motifs. This data predicts that motifs present upstream of the TA12015 gene can be recognised by the AP2 domain it encodes and the domains encoded by other AP2 factor genes that are up-regulated during differentiation to the merozoite.

The 'ACACACA' motif was previously found to be the most overrepresented motif throughout non-coding regions in *T. annulata* and *T. parva* (Guo and Silva, 2008), *Plasmodium* and *Toxoplasma* genes (Bohne et al., 1997; Guo and Silva, 2008, Young et al., 2008). However, Young et al. (2008) observed the 'NTGTGTGA' motif in a large group of genes expressed during the middle to later stages of the *Intraerythrocytic Developmental Cycle (IDC)* of *P. falciparum*. Analysis of upstream regions of *T. annulata* stage-regulated genes revealed enrichment of

this motif in those up-regulated during merogony, whereas a significant depletion ($p < 0.05$) of the motif was detected in macroschizont genes down-regulated during differentiation to the merozoite. A significant enrichment of the motif in genes up-regulated from merozoite to piroplasm stage was also obtained ($p < 0.05$). Taken together, these results suggest an association of this motif with stage differentiation to the merozoite/piroplasm, despite an occurrence of nearly 25% in the intergenic regions of the genome. This may point to a general type of function for factors that bind to the motif, the most obvious being modulation of chromatin status. Thus factors that bind to this motif may act as accessory factors for transcriptional regulation rather than as direct activators or repressors. The importance of accessory factors, such as HMG proteins, acting as hubs of nuclear function in higher eukaryotic cells is well known (Reeves et al., 2001).

A 'CACACA' motif was previously found in the upstream region of the merozoite surface antigen gene *Tams1* proximal to a motif 'TTTGTAGGG' (Shiels et al., 2000) that was specifically bound by a complex in parasite-enriched nuclear extracts of cells undergoing merogony. It was postulated that the 'CACACA' sequence was involved in stabilising the protein-DNA complex formation, as its deletion had a quantitative effect on factor binding. As this motif has also been found to be a binding site for some *Plasmodium* ApiAP2s it will be of importance to continue its analysis by use of experimental approaches to validate functionality of this putative DNA binding site in *T. annulata*. It is possible that an ApiAP2 factor binding to the 'CACACA' sequence may form a heterodimer complex or promote binding of a second factor and enhance *Tams1* gene expression during differentiation to the merozoite.

In summary, upstream regions of *T. annulata* genes with similar stage-regulated expression profiles show enrichment for motifs that could be targets for DNA binding proteins, indicating regulation by the same or related transcription factors. However, the results of this chapter also indicate that diversity in gene expression is probably associated with arrangements of distinct motifs that generate different combinations of DNA binding factors. While bioinformatic methods provide a starting point to explore regulatory motifs associated with stage differentiation, experimental work such as gel shift analysis is required to

validate recognition of putative motifs by DNA binding factors with emphasis on the up-regulated *T. annulata* ApiAP2 proteins identified in this study. Thus, while the data for TA13515 and TA11145 ApiAP2 genes (and their associated motifs) predict a function related to that of the *Plasmodium* orthologues, experimental validation of AP2 domain (and if possible factors in nuclear extracts) binding to predicted motifs is required. It was hoped that this could provide greater understanding of mechanisms involved in regulating differentiation to the merozoite in *Theileria* parasites.

Chapter 4

ApiAP2 genes as a candidate transcription factors for *T. annulata*

4 **ApiAP2 genes as candidate transcription factors for *T. annulata***

4.1 **Introduction**

After the completion of genome sequencing projects for *Apicomplexa* it became clear that there was a significant gap in understanding transcription regulation, as only a limited set of putative transcription factors were identified. However, in 2005 Balaji et al. described a group of proteins containing putative conserved AP2 DNA-binding domains (Apetala2 - integrase DNA binding domain) in *Plasmodium*. AP2 domain proteins were also found in other Apicomplexa species including *Theileria*, *Cryptosporidium* and *Toxoplasma* and are known nowadays as the Apicomplexan AP2 (ApiAP2) protein family (Campbell et al., 2010). ApiAP2 proteins are currently primary candidates for the major family of *T. annulata* transcription factors.

Full-length ApiAP2 proteins vary in size from a few hundred to several thousand amino acids. A single ApiAP2 domain has approximately 60AA and is built of three β -sheets and one α -helix. The domain also has a relatively long insert between the second and third β -sheet (Balaji et al., 2005), and this basic structure is highly conserved across all Apicomplexa species. ApiAP2 genes are not clustered on any particular chromosome or chromosomal region (De Silva et al., 2008). There is data suggesting that some ApiAP2 proteins may be processed during parasite development so that multiple forms may exist for individual proteins such as *Plasmodium* PFF0200c (PfSIP2) which has been shown to be proteolytically processed from the full length (230 kDa) form to a 50-60 kDa N-terminal segment containing the two ApiAP2 DNA-binding domains (De Silva et al., 2008).

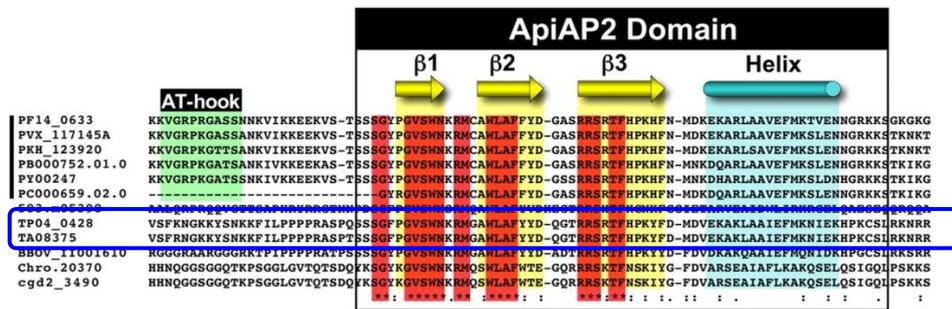


Figure 4.1. Alignment of the ApiAP2 domain from *P. falciparum* Pf140633 (amino acids 63-123) with five additional *Plasmodium spp.* and six apicomplexan (*T. gondii*, *T. parva*, *T. annulata*, *B. bovis*, *C. hominis* and *C. parvum*) orthologues (De Silva et al., 2008). Highlighted in dark blue box - *Theileria* ApiAP2 domains.

Conservation of residues likely to be involved in DNA binding are found in the three β -sheets (highlighted in red) with 12 residues highly conserved, based on at least 241 representatives of 285 diverse ApiAP2 domains tested by Balaji et al. (2005) (Figure 4.1). These residues appear to form key stabilising hydrophobic interactions and determine the backbone of the ApiAP2 domain. The crystal structure of the *P. falciparum* ApiAP2 domain from PF140633 which bound to its DNA motif has been also determined, revealing four important residues within the β -strand region that are directly in contact with the DNA. These four amino acids are highly conserved among all apicomplexan orthologues of PF140633 suggesting that DNA sequence specificity is well-conserved. β -sheets are stabilised by the alpha-helix which does not contact DNA (Allen et al., 1998; Lindner et al., 2010).

Complete understanding of the role of ApiAP2 proteins will require characterising possible stage-specific DNA interactions with other co-factor proteins. There is some evidence for this kind of interaction, although functional domains outside of the majority of ApiAP2 DNA - binding regions have yet to be identified. In 2010, Lindner et al. proposed a model of DNA-induced ApiAP2 dimerisation that would facilitate conformational rearrangements of the ApiAP2 protein or its partners. Binding of one ApiAP2 monomer to DNA induces a conformational change that recruits a second ApiAP2 domain, with the dimer forming a more stable interaction with the DNA. The ability of ApiAP2 proteins to form homo- and hetero-dimers is thought to increase the potential number of

target genes that could be differentially regulated by a small number of factors (Iyer et al., 2008; Campbell et al., 2010). Additional identified DNA binding proteins, such as Myb1, Myb2 or the high mobility group (HMGB) proteins will also likely contribute to the overall picture of transcriptional regulation. It can be predicted that the identification of ApiAP2 complexes will be a key step in describing transcription factor-based control of gene expression in *Plasmodium* and other Apicomplexa species (Painter et al., 2011).

Moreover, it has been previously demonstrated in plants that the *A. thaliana* CBF1, AP2 transcription factor interacts with GCN5, a histone acetyltransferase (HAT) that functions as a co-activator in transcriptional regulation (Stockinger, 2001) and these homo- and heterotypic interactions were also demonstrated for *Plasmodium* (LaCount et al., 2005). In *T. gondii*, the association of an ApiAP2 protein with GCN5 has also been successfully demonstrated (Painter et al., 2011). Recently GCN5b lysine acetyltransferase was proven to interact with *Toxoplasma* ApiAP2 factors and play a role in gene expression regulation and *T. gondii* proliferation (Wang et al., 2014). ApiAP2 are likely to function by recruiting histone-modifying and chromatin remodeling factors to promoter sites of their target genes. Depending on the type of factor bound by them, they might be involved in either transcription activation (by binding for example a member of GCN5-family lysine acetyltransferases) or repression (by recruiting CHD1 histone deacetylase) (Iyer et al., 2008).

The identification of the DNA-binding specificities of ApiAP2 proteins is very important for exploration of their role as transcriptional regulators during all stages of parasite development. Their striking differential expression in specific developmental stages of apicomplexans such as *Plasmodium* and *Toxoplasma* strongly suggest that they may also play an essential role in mediation of transcriptional regulation of *Theileria* stage specific genes. An analysis of the expression patterns of ApiAP2 genes in *P. falciparum* confirmed that different guilds of these transcription factors are specifically expressed in four major intraerythrocytic developmental stages - ring, trophozoite, early schizont and late schizont-merozoite stage (Balaji et al., 2005; Iyer et al., 2008). Microarray data has demonstrated that 21 of the 27 *P. falciparum* ApiAP2 genes are transcribed during the intraerythrocytic development cycle. Transcripts of

Pf13_0235 (Mikołajczak et al., 2008) and Pf14_0633 (Yuda et al., 2010) were found to be up-regulated in the sporozoite stage. Moreover, the *P. yoelli* orthologue of Pf13_0235 is highly expressed during the liver developmental phase (Tarun et al., 2008). In addition, very few ApiAP2 genes have been successfully knocked out, indicating that they are essential to blood-stage development and that the ApiAP2 proteins might maintain the progression of this developmental process. However, ApiAP2 factors may act also as suppressors of gene expression or have roles in regulating chromatin structure (Behnke et al., 2010). In *Plasmodium* Pff0200c (PfsIP2) domains were found to bind to SPE2 motif (GTGCAC) located upstream of var genes, and it was proposed that this particular *Plasmodium* ApiAP2 is involved in maintenance of heterochromatin more than acting as a regular specific transcription factor (Flueck et al., 2010).

An interesting example of a differentially expressed ApiAP2 is *Plasmodium* AP2-O protein (Pf11_0442; Pbanka_090590). AP2-O activates gene expression in the ookinete, which is a motile stage that invades the mosquito midgut. Similar to AP2-Sp (Yuda et al., 2010), the AP2-O gene is expressed in the preceding stage (gametocyte) and repressed, at the translational level by the DOZI (development of zygote inhibited) complex until ookinete formation (Mair et al., 2006). The amino acid sequences of the single ApiAP2 domain of this gene are almost identical among *Plasmodium* spp. orthologues, strongly suggesting that they mediate transcriptional regulation of stage specific genes across the *Plasmodium* genus (Yuda et al., 2009). Also the knockout of this gene generated an inability to undergo oocyst development (Mair et al., 2006). AP2-O requires two copies of a TAGCTA DNA motif, present in ookinete target genes, for high affinity binding (Yuda et al., 2009).

Interestingly, *Toxoplasma gondii* ApiAP2 genes are also cell cycle regulated and 24 ApiAP2 genes of *Toxoplasma* show peak expression times distributed throughout the tachyzoite cell cycle. This suggests a role in regulating progression of the tachyzoite cell cycle. In addition, surveys based on or using microarray (data) in *Toxoplasma* identified 11 ApiAP2 mRNAs induced during bradyzoite differentiation (Painter et al., 2011), indicating they may have roles in cyst formation. Radke et al. (2013) showed that the *Toxoplasma* AP2IX-9 protein, which binds to the CAGTGT DNA motif, is expressed at minimal level in

tachyzoites and fully expressed in the early bradyzoite stage nucleus. AP2IX-9 acts as repressor of bradyzoite genes; its over-expression significantly reduced tissue cyst formation, whereas loss of gene functionality caused an increase in the tissue cyst formation.

Other interesting motifs recognised by ApiAP2 factors that may have a role in *Theileria* stage differentiation are - the G-box motif bound by *P. falciparum* PF13_0235 (Campbell et al., 2010) and *C. parvum* Cgd8_810 (Oberstaller et al., 2013). This motif was found in upstream regions of *Plasmodium* ribosomal and heat shock genes (Millitello et al., 2004) and a heat shock/stress response has been implicated in differentiation to the merozoite in *T. annulata* (Shiels et al., 1992). In addition, motifs related to the CACACA motif (De Silva et al., 2008), also found in upstream regions of *T. annulata* genes (Guo and Silva, 2008 - see previous chapter) are of considerable interest. In *Plasmodium*, a group of four related ApiAP2 domains (Pfd0985w, Mal8P1.153 and Pf14_0533, and Pf13_002) recognise this type of sequence (see Figure 4.2), the first three of which are expressed during IDC (late trophozoite/early schizont stage; Campbell et al., 2010).

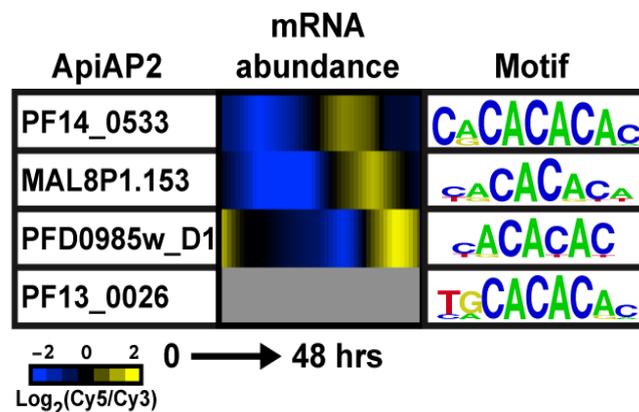


Figure 4.2. *Plasmodium* ApiAP2 proteins binding the CACACA motif. Three of these ApiAP2 factors are expressed in the late IDC stages (Campbell et al., 2010).

In addition to the primary DNA motif, ApiAP2 domains can have up to four secondary motifs that display changes to end or core nucleotides or differences in spacer distances, which could significantly increase the number of genes they bind to and the complexity of the regulatory network. For example, domain 2 of *P. falciparum* ApiAP2 PFD0985 is predicted to have three secondary motifs, in addition to the primary GTGTTACAC motif (Campbell et al., 2010). Also, the DNA-binding domain of *P. falciparum* AP2-Sp recognises several eight-base

sequences, beginning with TGCATG, present in the proximal promoter region of all known sporozoite-specific genes, but the specific binding properties vary across the different motifs (Yuda et al., 2010). Despite recognition of secondary motifs by individual ApiAP2 domains there is little overlapping binding specificity between different domains. Indeed, many upstream sequence elements have more than one ApiAP2 binding site suggesting combinatorial gene regulation. This may provide the diversity required to control a large number of genes using a small number of factors. Interestingly, in addition to regulating other genes, ApiAP2 factors can also form an interaction network with themselves by binding to the target motifs in their upstream regions (Campbell et al., 2010). The occurrence of target motifs in upstream regions of ApiAP2s and cascades of gene expression during developmental events suggests that auto-regulatory feedback involving ApiAP2 transcription factors is also likely to occur. Indeed it has been proposed for *P. falciparum* that AP2-G (Kafsack et al., 2014) and PFF0200c (De Silva et al., 2008) are auto-regulated, supporting the original hypothesis of a stochastic model of differentiation for apicomplexan parasites (Shiels et al., 1999).

Based on detailed studies performed in related Apicomplexa, the ApiAP2 proteins are excellent candidates for regulators of precisely timed and coordinated gene expression events associated with multistage differentiation steps in *T. annulata*. Importantly, identification of DNA sequence motifs bound by ApiAP2 factors allows for the prediction of putative target genes based on motif occurrence in upstream regions. Furthermore, these targets provide a good starting point for characterising a potential role for ApiAP2 proteins in stage differentiation. A better understanding of these transcription factors may facilitate investigation of strategies to disrupt interaction of ApiAP2 proteins with their target DNAs and inhibit parasite stage development. Investigating whether ApiAP2s are involved in the differentiation process of *T. annulata* may also shed further light on establishment of common regulatory processes across the Apicomplexa.

The main aim of the work presented in this chapter was to characterise selected putative stage-specific *T. annulata* transcriptional regulators and identify their associated nucleotide motifs. Based on previous data (see Chapter 2) the four

most up-regulated TaApiAP2 genes at the RNA level during differentiation from macroschizont to merozoite stage (TA13515, TA11145, TA12015 and TA16485) were chosen for further investigation. It was hoped that these TaApiAP2 domains could be shown to bind specific motifs recognized by their *Plasmodium* orthologues. Finally, it was a significant aim to test if any identified motifs could bind native nuclear factors and generate data that would implicate motif and factor in a mechanism regulating differentiation from the macroschizont to merozoite.

4.2 Material and Methods

4.2.1 ApiAP2 proteins: general characteristics and selected domain conservation analysis across *Theileria* and *Plasmodium* species

General characterisation of the 22 *T. annulata* sequences predicted to encode ApiAP2 proteins transcription factors was performed with data obtained from the PiroplasmaDB Genomic Resource, version 5.0 (<http://piroplasmadb.org>). Generated data included: position in genome, predicted function, size of gene and protein, ApiAP2 domain number and position in protein.

Five TaApiAP2 domains potentially binding to the CACACA type motif (TA11145, TA07100, TA07550, TA19920 and TA02615) based on their orthology with *P. falciparum* CACACA motif binding ApiAP2s (Balaji et al., 2005; Campbell et al., 2010) were aligned together with their *T. parva*, *T. orientalis* and *P. falciparum* orthologues. Domain boundaries were defined using the Pfam database (<http://pfam.sanger.ac.uk>). A phylogenetic tree was constructed with the use of Jalview (neighbor joining method using percentage identity) on the basis of a measure of similarity between each pair of sequences in the alignment. Branch length was estimated from the differences between these sequences.

4.2.2 ApiAP2 domain structure prediction

To predict protein structure and function, Phyre2 was used (Protein Homology/Analogy Recognition Engine; www.sbg.bio.ic.ac.uk/phyre2/index.cgi). Secondary and three-dimensional (3D) structure of selected ApiAP2 domains was predicted based on orthologue sequence comparison and experimentally determined structure. The Phyre2 server uses the Structural Classification of Protein (SCOP) (Murzin et al., 1995) and Protein Data Bank (Berman et al., 2000) databases as a library of known proteins and scans the query sequence with the use of PSI-Blast in order to construct the protein profile. Domain secondary structure was predicted within Phyre2 using three independent secondary structure prediction programs: Psi-Pred (McGuffin et al., 2000), SSPro (Pollastri et al., 2002) and JNet (Cole et al., 2008), with the output in the form of the

three-state prediction: alpha helix, beta strand and coil. The 3-D model of the ApiAP2 domains was finally constructed based on the protein homology profile and its secondary structure based on assumption that two homologues will share very similar structures. Generated by Phyre2, pdb files were then visualised by FirstGlance software (<http://bioinformatics.org/firstglance/fgij/>) to produce a final 3-D structure prediction of ApiAP2 domains.

4.2.3 Generation of selected *Theileria* ApiAP2 GST-fusion proteins

The P-gex glutathione S-transferase (GST) fusion protein system is a widely used method for high expression and fast purification of fusion proteins produced in *E. coli* cells. It has been used in many studies including DNA-protein interactions (De Silva et al., 2008; Campbell et al. 2010). Glutathione S-transferase (GST) is a 211 amino acid protein (26kDa) which is fused to the N-terminus of the recombinant protein. Fusion proteins are constructed by inserting a gene or gene fragment into selected cloning site of the pGex vector. Expression of the GST-fusion protein is then induced by isopropyl β -D thiogalactoside (IPTG). The pGex-X vector provides three translational reading frames beginning with the Eco R I restriction site (Figure 4.3). Use of two different restriction enzymes enables directional cloning of inserts into the vector. GST-fusion proteins are purified from bacterial lysates by affinity chromatography via enzyme-substrate binding using immobilized glutathione. GST-fusion protein is then eluted from the column using reduced glutathione. A detailed protocol for generation *T. annulata* GST-ApiAP2 fusion proteins is presented below. Recipes of buffers and solutions used to generate GST-fusion proteins can be found in Appendix (1.1).

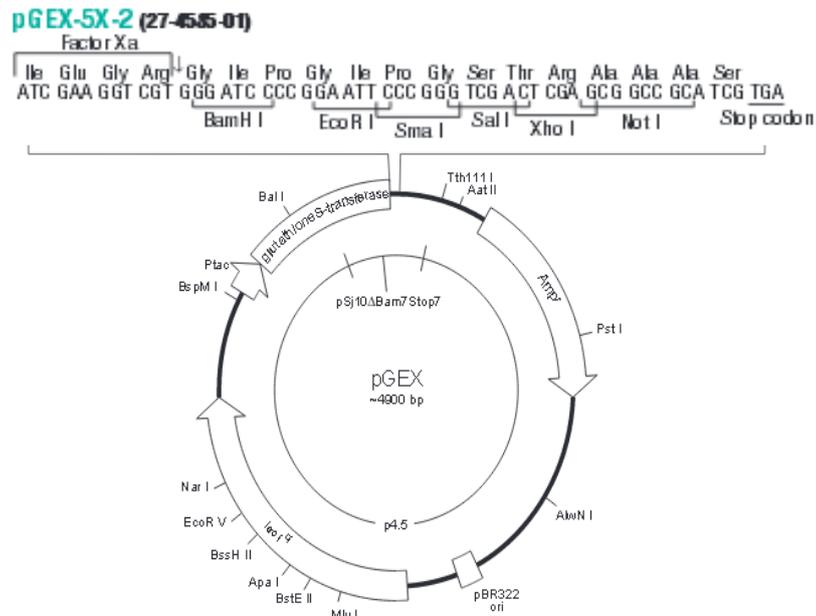


Figure 4.3. Map of the GST- fusion vectors showing the reading frames and main features.
 (http://brf.anu.edu.au/files/gst_fusion_handbook.pdf).

4.2.3.1 PCR amplification and cloning of ApiAP2 domains from *T. annulata* genomic DNA

TA13515, TA11145, TA16485 and TA12015 ApiAP2 domain boundaries were defined as in Balaji et al. (2005) and confirmed by Pfam domain search (<http://pfam.sanger.ac.uk/search>). Extensions of 10-20 nucleotides were included to make sure whole domain was covered and primers designed to create N-terminal GST-fusion constructs by cloning. Primers contained EcoRI and XhoI restriction sites - forward with EcoRI and reverse with XhoI (see Appendix 1.8), and manufactured by Eurofins MWG Operon, Germany. Primers were dissolved in Ultra pure sterile H₂O to 100pmol/μl stock. PCR for each gene was performed with a proof-reading activity Polymerase (Pfu), in triplicate. *T. annulata* DNA (strain Ankara) isolated from piroplasms purified from the blood of an infected calf (kindly donated by Dr Laetitia Lempereur, Glasgow) was used as template. PCR amplification of the four ApiAP2 domains was then performed (see Appendix 1.6 for reaction components and conditions). 5 μl of each PCR sample was then run on 1.5% agarose TAE gel to determine fragment size. A QiaQuick PCR Purification Kit (Qiagen, 28104) was then used, according to the manufacturer's instructions

(http://2012.igem.org/wiki/images/a/a3/QIAquick_PCR-purification.pdf), to remove free oligos and dNTPs. Purified DNA amplicons were eluted using 30µl elution buffer.

4.2.3.2 Construction of recombinant plasmids

28µl of DNA representing individual ApiAP2 domain DNA was double digested with 2µl EcoRI and 2µl XhoI together with 4,5µl of 10xReact2 buffer and 4µl of ddH₂O. Following incubation for 2h at 37°C, enzymes were inactivated by incubating samples in a 65°C water bath for 10 min. Samples were then run on a 1.5% agarose gel and DNA fragments excised, under long wave UV visualisation. Samples were placed in 1.5 ml eppendorf tubes and weighed. QiaQuick gel extraction for small fragments was then performed, according to the supplied protocol(www.sites.bio.indiana.edu/~chenlab/protocol_files/agarose_gel_extraction.pdf). Samples were eluted in 30µl of elution buffer and fragment concentration estimated by running on 1.5% agarose gels relative to size markers (10, 20, 50 and 100ng - 10µl per lane). pGex5x-2 vector (Sigma-Aldrich, GE28-9545-54) DNA was also cut with EcoRI and XhoI restriction enzymes in x1 One-Phor-All Buffer PLUS according to the GST Gene Fusion System Handbook instructions (http://brf.anu.edu.au/files/gst_fusion_handbook.pdf) and stored at a concentration of 40ng/µl. A ligation calculator (www.insilico.uni-duesseldorf.de/Lig_Input.html) was then used to calculate amount of vector and insert needed, and a vector to insert ratio of 1 to 3 was used. Vector and insert were mixed with 2µl of 5x ligase buffer, 1µl of ligase and water added to a final volume of 10µl. Reactions were left O/N at 15°C and 90µl of ddH₂O added the next day. Ligated plasmids DNA were stored frozen in -20°C or directly followed by bacterial transformation.

4.2.3.3 Bacterial transformation

Aliquots of competent XL-1 Blue cells (Stratagene, 200249) were split into 2x100µl following the bacterial transformation protocol of the supplier (www.genomics.agilent.com/files/Manual/200249.pdf). 10µl of each ligation reaction was added to competent cells and 0.9ml SOC medium. 100-300µl of

each transformation was then spread on LB/ampicillin (50 µg/ml) plates and incubated O/N at 37°C. Visible colonies were counted the next day. Colonies were picked into separate universal tubes containing 5ml of LB with antibiotic (final concentration 100µg/ml of Ampicillin). They were grown O/N with shaking at 37°C and plasmids were isolated the next day using the PureLink Quick Plasmid Miniprep Kit (Invitrogen™) according to manufacturer's instructions (www.tools.lifetechnologies.com/content/sfs/manuals/purelink_quick_plasmid_qrc.pdf). Plasmid DNA was then digested with EcoRI and XhoI enzymes to check for inserts of the correct size. Alternatively, PCR was performed (using the ApiAP2 domain specific primers) to check if the insert could be amplified. Stocks of bacteria transformed with recombinant plasmids were stored at -80°C (150µl of autoclaved glycerol + 850µl of transfected *E. coli*). Plasmid DNA was stored at -20°C. DNA samples were sent for sequencing (Eurofins MWG Operon, Germany) with commercial pGex5x-2 sequencing primers to validate the construct representing the selected ApiAP2 domain based on the sequence from *T. annulata* Gene DB (www.genedb.org/Homepage/Tannulata).

4.2.3.4 Test induction of GST fusion proteins

For expression of fusion protein, selected pGex constructs were re-transfected in BL21 Codon Plus (DE3)-RIL (Stratagen™) competent cells, according to the manufacturer's instructions (http://www.med.unc.edu/pharm/sondeklab/Lab%20Resources/manuals/codon_plus_manual.pdf), spread on ampicillin plates and grown O/N at 37°C. A colony from each transfection was transferred into 10ml LB medium with antibiotic and grown O/N at 37°C in a shaking incubator. An experiment to test fusion protein expression and solubility was set up. Two vials for each construct were prepared, each with 10ml of LB medium/antibiotic and 1ml (1/10 dilution) of fresh overnight culture. Cells were grown for 2h at 37°C with shaking and IPTG added (final concentration of 0.2mM) to one of the vials. Cultures were grown for a further 4h at 26°C and then transferred to 15ml centrifuge tubes and centrifuged at 5000rpm (1956xg) for 15 min. Cell pellets were then washed X2 with 5ml of PBS, centrifuged (as above), drained carefully and stored at -20°C. The next day pellets were thawed and resuspended in 200µl of PBS containing fresh 1mg/ml lysozyme. Cells were then incubated on ice for 30min and

sonicated 4 times, 6 sec burst at amplitude 7 with 10 sec cooling intervals between each burst. 20µl of 10% TritonX-100 was added to the tube, mixed and tubes incubated on ice for another 20min. Samples were then centrifuged at 13000rpm (13225xg) for 15min at 4°C. Supernatant containing the soluble fraction was transferred into a fresh tube and 200µl 2xSDS sample buffer was added. Pellets containing the insoluble fraction were resuspended in 200µl+200µl 2x SDS sample buffer and boiled for 5min. 10µl of each sample (supernatant and pellet, un-induced and induced for each clone) was run on a 10% SDS-page gel at 120V for 2h and the gel stained with Coomassie Blue (1h) and destained (O/N).

4.2.3.5 Enhancing solubility of expressed recombinant proteins in *E. coli*

One of the disadvantages of using *E. coli* cells for an expression system is the formation of insoluble aggregates (known also as inclusion bodies) of folding protein intermediates. This is caused by the rate of protein translation exceeding the capacity of cells to fold newly synthesised protein correctly (Kiefhaber et al., 1991). Decreasing the rate of protein production is one of the strategies used to overcome this problem. Unfortunately, several methods used in this study (applying a lower temperature of induction, induction at a low IPTG concentration) failed to increase expression of soluble GST-TA12015D. A novel method of enhancing solubility of the expressed GST-fusion proteins in *E. coli* (Ghosh et al., 2004) by inducing GST-TA12015D expression in the presence of the dipeptide glycylglycine (Gly-Gly - Sigma, G1002). GST-TA12015D protein expression was performed adding different concentrations (0, 50, 200, 500mM and 1M) of Gly-Gly to the TB medium.

4.2.3.6 Protein expression for large-scale production

Two 250ml flasks of TB (Terrific Broth) with ampicillin (with addition of 500nM of Gly-Gly for GST-TA12015D expression) were inoculated with BL21 competent cells transformed with AP2-PGex constructs (1/10 dilution of fresh over-night culture). 0.2% (w/v) glucose was added to prevent inclusion body formation. Cultures were grown at 37°C for 1h (until OD600 reached 0.2), and the temperature reduced to 30°C for another 1h (till OD600 of 0.6). 0.2mM IPTG was then added to one of the flasks to induce protein expression and cultures grown for another 4h at 26°C. Cultures were transferred to pre-weighed 50ml tubes,

and spun at 6000xg for 10 min (3000rpm for 20min), the resulting supernatant was decanted and the pellet washed 1x in cold PBS. Drained pellets were weighed and stored at -20°C.

10ml of fresh GST buffer with 100µl of 100mM PMSF, 50µl of 10% TritonX-100, 1mg/ml lysozyme and protease inhibitor cocktail (Sigma Aldrich®) was added to thawed samples. Samples were incubated for 30min at 4°C on a nutator at 20rpm and frozen for 30min (to overnight) at -80°C. Cells were then thawed and sonicated using the 6mm probe in a cut-off 50ml falcon tube sitting in ice water at amplitude 7 for 15s, with 30s cooling for 7 cycles. Cell debris was spun out at 4000rpm (3094xg) for 30min and the supernatant transferred to another tube and kept on ice.

4.2.3.7 GST- fusion protein purification

Glutathione sepharose affinity beads (Sigma-Aldrich®, GE17-0756-01) were prepared according to the manufacturer's instructions (http://www.gelifesciences.com/gehcls_images/GELS/Related%20Content/Files/1314807262343/litdoc18115758_20140113001413.pdf) and added to 100ml of prepared supernatant and kept at 4°C for 30min. The GST column was washed X2 with 2ml of cold PBS + 0.1% TritonX-100 and then samples run through the column. The column was washed again with 10x5ml PBS+Triton, and each wash collected into separate tubes for further testing. GST-fusion proteins were eluted with 10x500µl of glutathione elution buffer, with each elution fraction collected into separate tubes. As most protein will be in fractions 2-7, 10µl of each fraction was removed for analysis on SDS PAGE gel before pooling fractions with the clearest profile. Additionally, flow-through samples and the 10th wash sample were run on the gel to make sure the GST-fusion protein was eluted completely from the column. Protein concentration were established by use of Coomassie Plus - The Better Bradford™ Assay Reagent, (Pierce Biotechnology, 23238) according to manufacturer's instructions (www.wolfson.huji.ac.il/purification/PDF/Protein_Quantification/PIERCE_Coomassie_KIT.pdf).

GST-fusion proteins were concentrated using Amicon[®] Ultra-15 Centrifugal Filter Units with an Ultracel-3 membrane (Millipore, UFC900308: [www.millipore.com/userguides/files/centrifugal/\\$file/PR03520TR_RevA_English.pdf](http://www.millipore.com/userguides/files/centrifugal/$file/PR03520TR_RevA_English.pdf)). Eluted proteins were stored at -80 °C in 25-50µl aliquots, at concentration 1mg/ml.

4.2.4 Chemiluminescent Electrophoretic Mobility Shift Assay (EMSA)

Electrophoretic Mobility Shift Assay (EMSA) has become a technique of choice for studying qualitative and quantitative protein-nucleic acid interactions and determining potential gene regulation (Garner and Revzin, 1986; Fried, 1989). The Thermo Scientific LightShift[®] Chemiluminescent Electrophoretic Mobility Shift Assay is a fast and easy method for detecting DNA and protein interactions via fluorescence detection. This technique is based on DNA-protein complex migration time in comparison to free non-bound DNA probe (Figure 4.4). Purified fusion protein or nuclear extract are incubated with biotin-labeled DNA and then run on a native polyacrylamide gel and transferred to a nylon membrane. The biotin-labeled DNA is detected using Streptavidin-Horseradish Peroxidase Conjugate and a Chemiluminescent substrate. Poly (dI•dC) (Poly(deoxyinosinic-deoxycytidylic) acid) used in the reaction mix is a DNA competitor used for blocking non-specific binding of proteins to the labeled probe. Unlabelled-specific competitor may be added in excess to remove the band shift and confirm DNA binding motif specificity. Use of a mutated biotin-labeled oligo provides an easy way to determine specificity of the motif sequence. The control EBNA System was used according to the manufacturer`s instructions (www.piercenet.com/instructions/2160919.pdf).

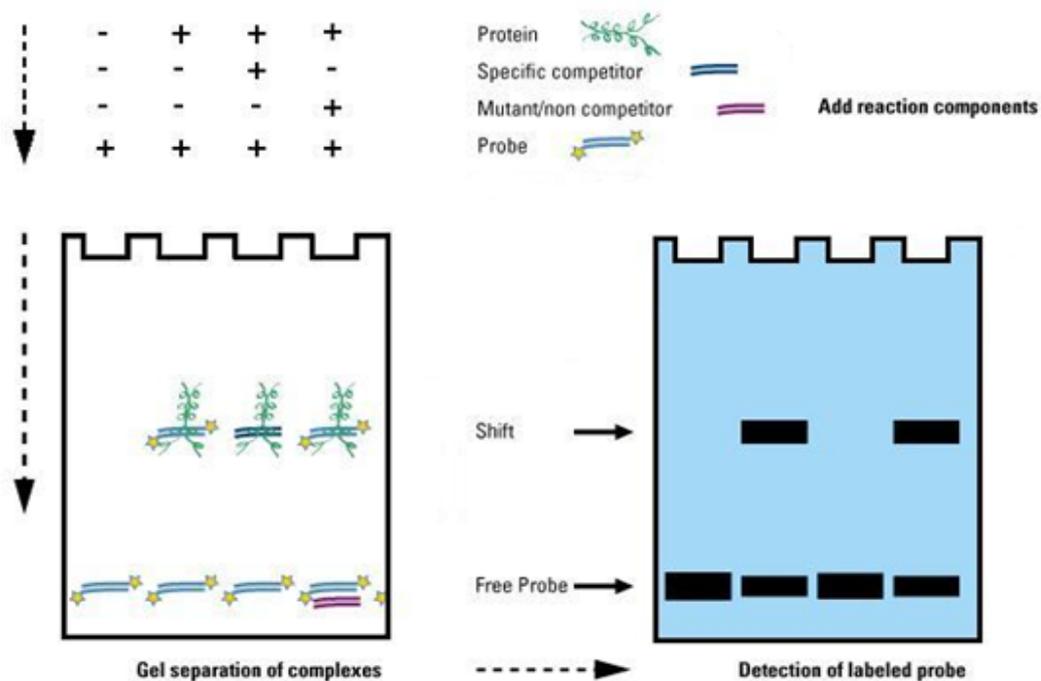


Figure 4.4. Electrophoretic mobility shift assay (EMSA). Binding reaction complexes are separated by electrophoresis and then transferred onto a membrane to detect the biotinylated probe. Addition of specific competitor completely eliminates the protein-DNA shift (www.piercenet.com/method/gel-shift-assays-emsa).

All EMSA reaction components were added in the specific order presented in Table 4.1. for GST-fusion protein reactions and Table 4.2., for reactions with Parasite Nuclear Extracts (PNE) derived from the D7 infected cell line (donated by Prof. B. Shiels and Dr J. Kinnaird, Glasgow University). Parasite enriched fractions were obtained based on the method of Shiels et al. (2000) but using the NE-PER™ Nuclear and Cytoplasmic Extraction Reagent kit, following the supplier's instructions (Thermo Scientific™, 78835; www.piercenet.com/instructions/2160872.pdf).

Single-stranded HPLC purified 5' biotinylated oligonucleotides containing an ApiAP2 target or mutated motif (synthesized by Eurofins Genomics, Germany) were diluted in ddH₂O to 100 pmol/μl. Equal volumes of both labelled and complementary unlabelled oligonucleotides (at equimolar concentrations) were mixed with 10x annealing buffer (1ml Tris-HCl, 1ml 100mM; pH 8, NaCl (2M) and 8ml ddH₂O) and annealed using a thermocycler following the protocol of Thermo scientific (www.piercenet.com/files/TR0045-Anneal-oligos.pdf). Annealed oligonucleotides were diluted to 1pmol/μl for non-labelled and 20 fmol/μl for biotinylated probes. All oligos used in this study can be found in Appendix (1.9).

A 4% polyacrylamide gel was prepared (see Appendix - 1.1) and run at 100V, at 4°C. Free and bound probes were transferred to Biodyne® Precut Nylon Membrane (Thermo Scientific, 77015), cross-linked to the membrane at 120mJ/cm² using UV-light and detection performed with the Chemiluminescent Nucleid Acid Detection Module (Thermo Scientific, 89880; www.regulatorylogic.files.wordpress.com/2012/01/chemiluminescent-nucleic-acid-detection-module-thermo-scientific-898801.pdf).

DNA-PROTEIN	
Ultrapure water	to 20µl
10x binding buffer (kit)	2 µl
1µg/µl Poly (dIdC) (kit)	1 µl
50% glycerol (kit)	1 µl
100mM MgCl₂ (kit)	1 µl
1% NP40 (kit)	1 µl
Unlabelled oligo (1pmol/µl)	x µl
Protein (0.7-1mg/µl)	2 µl
Biotinylated oligo (20fmol/µl)	1 µl

Table 4.1. EMSA reaction components to study ApiAP2 domain fusion protein-DNA interactions.

DNA-PNE	
Ultrapure water	to 20µl
10x binding buffer (kit)	2 µl
1µg/µl Poly (dIdC) (kit)	1 µl
50% glycerol (kit)	1 µl
100mM MgCl₂ (kit)	1 µl
1% NP40 (kit)	1 µl
EDTA	1 µl
Unlabelled oligo (1pmol/µl)	x µl
PNE	5 µl
Biotinylated oligo (40fmol/µl)	2 µl

Table 4.2. EMSA reaction components to study parasite nuclear extract-DNA interactions.

4.2.4.1 EMSA with Polyamide inhibitors

EMSA was also performed in the presence of polyamide synthetic ligands (PAs) containing N-methylimidazole and N-methylpyrrole amino acids, which by binding to specific DNA sequences compete out the binding of transcription factors and modulate gene expression. Polyamide molecules bind DNA in the minor groove by forming sequence specific side-by-side aromatic acid pairing complexes: Imidazole (Im) opposite Pyrrole (Py) targets a G•C base pair, Py/Im targets C•G and Py/Py targets T•A and A•T with affinities comparable to those of natural DNA-binding transcription factors.

EMSA was performed with GST-TA13515D fusion protein and an ISS-15 inhibitor specific for the GTGTAC sequence (target motif of PfAP2-G), and as a negative control an ISS-33 inhibitor specific for the TAGCTA sequence (target motif for PfAP2-0) was used. Both inhibitors were kindly donated by Prof. A. Waters (Glasgow University). PA-EMSA reaction components were added in the order presented in Table 4.3. and the assay performed as above.

DNA-PROTEIN- INHIBITOR	
Ultrapure water	to 20µl
10x binding buffer (kit)	2 µl
1µg/µl Poly (dIdC) (kit)	1 µl
50% glycerol (kit)	1 µl
100mM MgCl₂ (kit)	1 µl
1% NP40 (kit)	1 µl
Inhibitor (x nM)	x µl
Protein (0,7-1mg/µl)	1 µl
Biotinylated oligo (40fmol/µl)	2 µl

Table 4.3. EMSA reaction components to study protein-DNA interactions in the presence of specific PA inhibitors.

4.2.5 Comparison of expression data from the D7 cell line and the D7B12 cell line attenuated for merogony

The D7 infected cloned cell line differentiates from the macroschizont to merozoite stage at 41°C, whereas in the cell line D7B12, derived from recloning D7, the parasite is severely attenuated for the differentiation step (Shiels et al., 1994). To compare expression patterns of the ApiAP2 encoding TA11145 and TA07100 genes between these two cell lines, qRT-PCR was performed for Day 0 (37°C), Day 4 and Day 7 (41°C) RNA samples derived from D7 and D7B12 (as described in Chapter 2). The list of specific primers designed for qRT-PCR is given in Appendix (1.7). qRT-PCR was performed as described in section 2.2.5.3 and data analysed by MxPro v4.10 software, as described in section 2.2.5.4. The relative quantity values were normalised to the housekeeping gene Hsp70 and fold change calculated relative to the calibrator condition, Day 0 - macroschizont infected cells. Generated graphs represent normalised mean values ± standard error of the mean (SEM). Statistical analysis was performed using a two-tailed Student's t-test in the Excel program. The difference between expression levels in D7 and D7B12 cell lines was considered to be significant if the p-value was less than 0.05.

4.2.6 Immunofluorescent Antibody Test (IFAT)

The Indirect Immunofluorescent Antibody Test (IFAT) is a two-step method for routine sero-diagnosis. In this study, an IFAT test of D7 cells undergoing a time-course differentiation was performed to assess whether changes observed by qPCR are reflecting modifications in antigenic profile associated with merogony. Two antisera were raised commercially (Scottish National Blood Transfusion Service) by immunization of host animals (rabbits) with antigens (TA13515D and TA11145D GST-fusion proteins). Pre-Immune sera were obtained to use as a negative control in IFAT. The obtained immune sera were run through a GST-column (see section 4.2.6.7) to remove antibodies that bound to GST.

D7 cells were grown to $\sim 3-8 \times 10^5 \text{ ml}^{-1}$ and 3×10^4 cells cytopun on to a glass slide at 1500RPM for 5 min (Shandon cytopspin 2 centrifuge). Slides were incubated in ice cold 3.7% p-formaldehyde in PBS for 30 min, rinsed briefly 3x in 1x PBS and

then permeabilised in pre-chilled methanol at -20°C for 10 min. After permeabilisation slides were washed 3 times in 1xPBS for ~1 min each wash. Drained slides were arranged in a wet box and RPMI-20% foetal calf serum (culture medium) was applied to block unspecific binding. 30 μl of anti-serum diluted at 1/100, 1/200 and 1/400 in culture medium was applied. Slides were then incubated in a wet box at room temperature for 1h and then washed 3 x in 1xPBS for 1 min each. Slides were drained and then incubated with 1/200 dilution of AlexaFluor 488 Anty-Rabbit Ig (Invitrogen) conjugated 2 $^{\circ}$ antibody in culture medium for 30 min. Slides were washed 3 x in 1xPBS 1min each, drained and counter-stained in 0.2% Evans blue, 1x PBS. Slides were then rinsed 2x in PBS, allowed to dry and mounted in 10 μl of mounting medium (50% glycerol, 4, 6-diamindino-2-phenylindole (DAPI) at 1 $\mu\text{g}/\text{ml}$ and an anti-fade DABCO) with a cover slip. Images were acquired using an Olympus BX60 microscope, a SPOT camera and SPOTTM Advanced image software (Version Mac: 4.6.1.26; Diagnostic Instruments, Inc).

4.3 Results

4.3.1 *T. annulata* ApiAP2 family, general characteristics

The *T. annulata* ApiAP2 family consists of 22 proteins (Table 4.4), the majority of which are annotated in Gene DB as hypothetical proteins (TA06995 is denoted as putative transcriptional adaptor - ADA2 homologue). However, due to the presence of the AP2 domain they are predicted to bind to specific DNA sequences, and could be involved in regulation of transcription. All the ApiAP2 genes are spread throughout the genome of *T. annulata* and are located in each of the 4 chromosomes, with the open reading frame encoded on either DNA strand.

Most *T. annulata* ApiAP2 proteins possess only one DNA-binding AP2 domain. Two proteins (TA07550 and TA20595) have two AP2 domains and one predicted protein - TA05055 possesses three AP2 domains. Very few of the ApiAP2 proteins have additional known conserved domains - TA07550 has a cytoadhesin-P30 domain thought to be important in cytoadherence and virulence; TA16485 and TA09965 possess the ACDC (AP2-coincident C-terminal) domain of unknown function, found previously at the C-terminus of other apicomplexan proteins containing the ApiAP2 domain. TA18095 encodes a clathrin adapter complex-related domain. Analysis of TA06995 by NCBI Conserved Domain Search revealed three additional domains in addition to the independent ApiAP2: a ZZ_ADA2 domain (zinc finger, coordinating two zinc ions and likely to participate in ligand binding or molecular scaffolding); SANT complex (SWI3, ADA2, N-CoR and TFIIB') DNA-binding domains that are often found in regulatory transcriptional repressor complexes; and a histone acetyltransferase complex SAGA/ADA, involved in regulation of chromatin structure and dynamics.

Gene ID	Annotation	Chromosome	Gene Strand	Exons	Transcript length	Protein length	AP2 domains	Other domains	AP2 position in protein	Predicted GO function	Predicted GO process
TA13515	hypothetical protein	2	reverse	1	1815	604	1	0	390 - 439	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA13395	hypothetical protein	2	forward	3	747	248	1	0	100 - 158	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA12015	hypothetical protein	2	reverse	2	855	284	1	0	66 - 127	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA11665	hypothetical protein	2	reverse	1	2355	784	1	0	179 - 237	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA02615	hypothetical protein	3	forward	1	903	300	1	0	227 - 285	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA04145	hypothetical protein	3	reverse	3	3477	1158	1	0	969 - 1029	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA04435	hypothetical protein	3	forward	5	1113	370	1	0	243 - 301	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA05055	hypothetical protein	3	forward	1	2256	751	3	0	386 - 445, 456 - 516, 699 - 751	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA17415	hypothetical protein	3	reverse	4	582	193	1	0	102 - 159	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA07100	hypothetical protein	4	reverse	1	1278	425	1	0	332 - 381	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent

Gene ID	Annotation	Chromosome	Gene Strand	Exons	Transcript length	Protein length	AP2 domains	Other domains	AP2 position in protein	Predicted GO function	Predicted GO process
TA08375	hypothetical protein	4	reverse	2	1230	409	1	0	66 - 126	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA10940	hypothetical protein	4	reverse	1	1191	396	1	0	227 - 282	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA09965	hypothetical protein	4	reverse	1	1203	400	1	1	210 - 268	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA11145	hypothetical protein	4	reverse	1	1737	578	1	0	518 - 576	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA19920	hypothetical protein	1	forward	1	2493	830	1	0	384 - 432	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA20595	hypothetical protein	1	forward	2	2157	718	2	0	201 - 295, 506 - 564	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA06995	transcriptional adaptor (ADA2 homologue), putative	1	forward	8	3441	1146	1	3	1093 - 1146	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA16105	hypothetical protein	1	reverse	1	867	288	1	0	222 - 278	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA16485	hypothetical protein	1	reverse	1	1662	553	1	1	280 - 332	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA18095	clathrin adapter complex-related protein, putative	3	forward	11	1113	370	1	1	268 - 322	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent
TA16535	hypothetical protein	1	reverse	3	1773	590	1	0	381 - 441	sequence-specific DNA binding transcription factor activity	regulation of transcription, DNA-dependent

Table 4.4. Characterisation of TaApiAP2 gene family. Highlighted in yellow - four most up-regulated from macroschizont to merozoite ApiAP2s.

4.3.2 ApiAP2 proteins structure prediction

Protein structure prediction via Phyre2 homology modelling is based on the assumption that two or more homologues proteins share very similar structure, which is evolutionary conserved. The accuracy of the protein structure prediction depends mainly on sequence similarity between the query sequence and template deposited in the database. As ApiAP2 protein conservation applies only to the ApiAP2 domain, a low overall sequence identity percentage was anticipated. However, at high confidence match (>90% confidence), even with a sequence identity of ~20%, the overall fold of the protein model is expected to be correct, and its central core to be accurate (Kelley and Sternberg, 2009). *T. annulata* ApiAP2 proteins (TA13515, TA11145, TA16485 and TA12015) showed similarity with a 2.2a crystal structure of the ApiAP2 domain of Pf14_0633 from *P. falciparum*, bound as a domain-swapped dimer to its cognate DNA; with a confidence of 98.5% (for TA11145) with 20% protein identity (Figure 4.6), 97.12% (for TA13515) with 27% protein identity (Figure 4.5), 61.6% (for TA16485) with 24% of protein identity (Figure 4.7) and 98.1% (for TA12015) with 22% identity (Figure 4.8). All proteins showed high conservation of the ApiAP2 domain. Additionally, part of the TA13515 protein showed statistically significant similarity to a beta-beta-alpha zinc finger domain with a confidence of 81.2% and 26% identity (Figure 4.5). According to the Phyre2 prediction TA16485 showed similarity to the transmembrane protein (colicin ia) with 92.3% confidence and 8% identity (Figure 4.7). However, 59% of the protein sequence was predicted disordered and these regions cannot be meaningfully structurally predicted.

Secondary structure prediction of TaApiAP2s (Figure 4.9-12) was built based on alignment with a known secondary structure template. The SS confidence line was added as a representation of the prediction confidence level: with red representing high confidence and blue low confidence. The disorder line represents prediction of regions that lack a fixed tertiary structure and these are indicated by question marks (?). It was observed that the weakest regions of secondary structure prediction corresponded to a strong prediction of disorder. 3D structure prediction of all analysed TaApiAP2s showed the presence of the three anti-parallel beta-strands that (have been shown) are predicted to be in

contact with target DNA sequence and the alpha-helix which stabilizes strands but does not have contact with DNA (Figure 4.13). These results support postulation that the TaApiAP2 domains function to bind specific DNA motifs in a similar manner to that described for related apicomplexan AP2 factors.

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c3igmA	 Alignment		97.1	27	PDB header: transcription/dna Chain: A; PDB Molecule: pf14_0633 protein; PDBTitle: a 2.2a crystal structure of the ap2 domain of pf14_0633 from p.2 falciparum, bound as a domain-swapped dimer to its cognate dna
2	d2ct5a1	 Alignment		81.2	26	Fold: beta-beta-alpha zinc fingers Superfamily: beta-beta-alpha zinc fingers Family: BED zinc finger
3	c2djrA	 Alignment		56.4	25	PDB header: metal binding protein Chain: A; PDB Molecule: zinc finger bed domain-containing protein 2; PDBTitle: solution structures of the c2h2 type zinc finger domain of 2 human zinc finger bed domain containing protein 2

Figure 4.5. View of the top three hits of the TA13515 fold recognition results, including images of the protein models produced. Matches are colour-coded for confidence value with a red background high confidence (97.2% probability) prediction.

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c3igmA	 Alignment		98.5	20	PDB header: transcription/dna Chain: A; PDB Molecule: pf14_0633 protein; PDBTitle: a 2.2a crystal structure of the ap2 domain of pf14_0633 from p.2 falciparum, bound as a domain-swapped dimer to its cognate dna
2	c2lurA	 Alignment		41.6	47	PDB header: plant protein Chain: A; PDB Molecule: kalata; PDBTitle: nmr solution structure of kb1[ghrw;23-28]
3	c2l23A	 Alignment		33.5	48	PDB header: transcription Chain: A; PDB Molecule: mediator of rna polymerase ii transcription subunit 25; PDBTitle: nmr structure of the acid (activator interacting domain) of the human2 mediator med25 protein

Figure 4.6 View of the top three hits of the TA11145 fold recognition results, including images of the protein models produced. Matches are colour-coded for confidence value with a red background high confidence (98.5% probability) prediction.

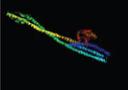
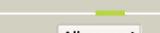
#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c1cia_A	 Alignment		92.3	8	PDB header: transmembrane protein Chain: A; PDB Molecule: colicin Ia; PDBTitle: colicin Ia
2	c3igmA	 Alignment		61.6	24	PDB header: transcription/dna Chain: A; PDB Molecule: pf14_0633 protein; PDBTitle: a 2.2a crystal structure of the ap2 domain of pf14_0633 from p.2 falciparum, bound as a domain-swapped dimer to its cognate dna
3	d1x9fc	 Alignment		55.5	16	Fold: Globin-like Superfamily: Globin-like Family: Globins

Figure 4.7. View of the top three hits of the TA16485 fold recognition results, including images of the protein models produced. Matches are colour-coded for confidence value with a red background high confidence (61.6% probability) prediction.

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	c3igmA	 Alignment		98.1	22	PDB header: transcription/dna Chain: A; PDB Molecule: pf14_0633 protein; PDBTitle: a 2.2a crystal structure of the ap2 domain of pf14_0633 from p.2 falciparum, bound as a domain-swapped dimer to its cognate dna
2	c4gwpB	 Alignment		49.7	23	PDB header: transcription Chain: B; PDB Molecule: mediator of rna polymerase ii transcription subunit 17; PDBTitle: structure of the mediator head module from s. cerevisiae
3	c4h63Q	 Alignment		32.4	17	PDB header: transcription Chain: Q; PDB Molecule: mediator of rna polymerase ii transcription subunit 17; PDBTitle: structure of the schizosaccharomyces pombe mediator head module

Figure 4.8. View of the top three hits of the TA12015 fold recognition results, including images of the protein models produced. Matches are colour-coded for confidence value with a red background high confidence (98.1% probability) prediction.

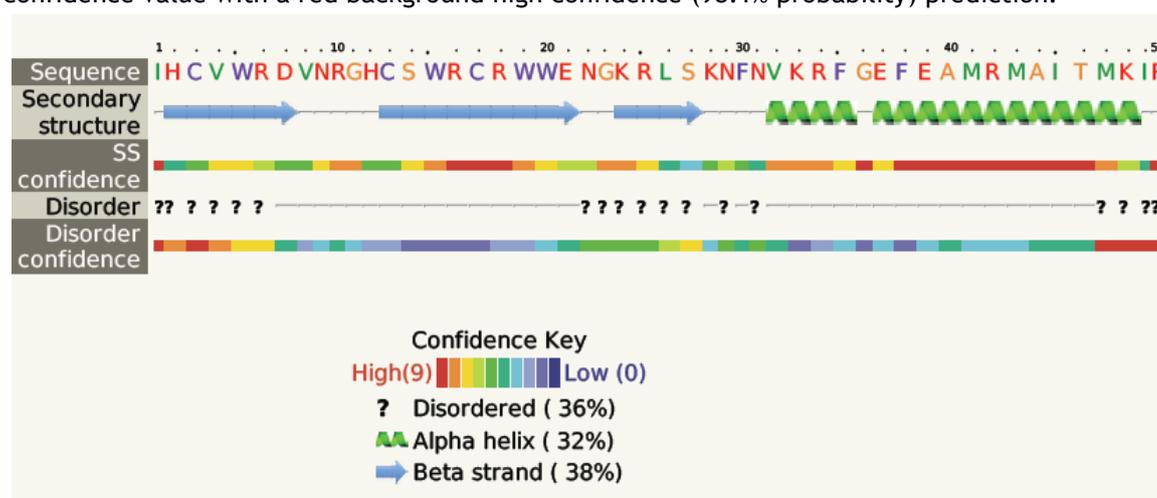


Figure 4.9. Secondary structure prediction of the TA13515 ApiAP2 domain. Blue arrows represent β -strands, green helices indicate α -helices and lines indicate coil state. Disordered regions are indicated by question marks (?). Prediction confidence levels are colour-coded with red representing high confidence and blue low confidence.

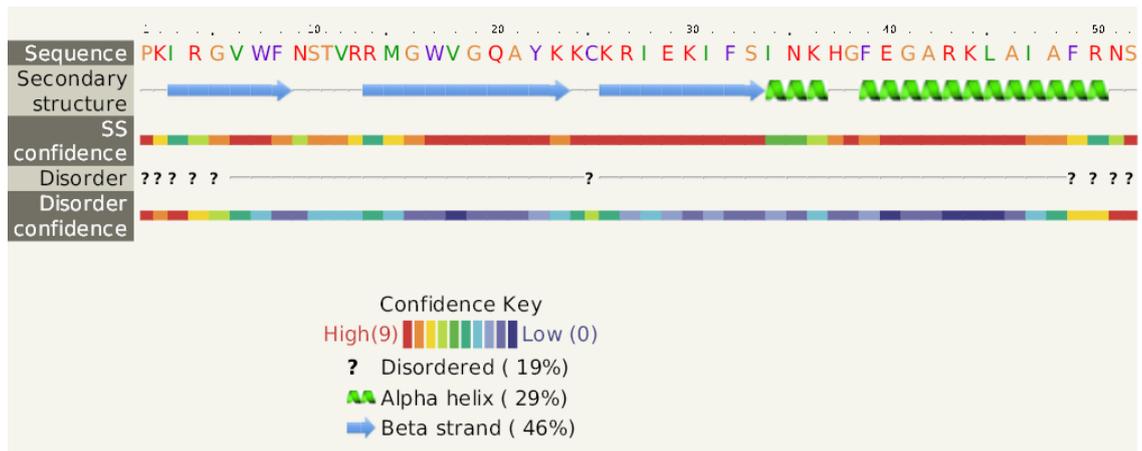


Figure 4.10. Secondary structure prediction of the TA16485 ApiAP2 domain. Blue arrows represent β -strands, green helices indicate α -helices and lines indicate coil state. Disordered regions are indicated by question marks (?). Prediction confidence levels are colour-coded with red representing high confidence and blue low confidence.

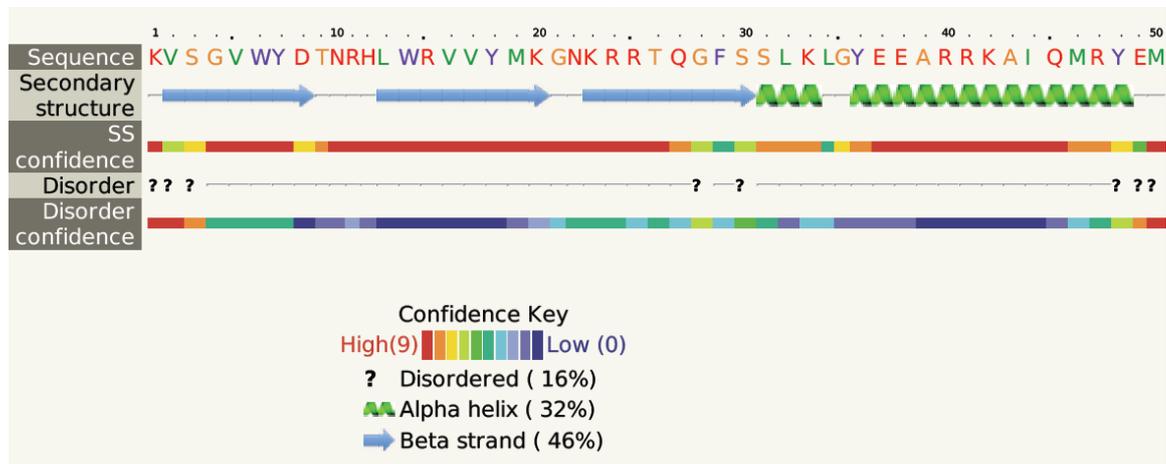


Figure 4.11. Secondary structure prediction of the TA11145 ApiAP2 domain. Blue arrows represent β -strands, green helices indicate α -helices and lines indicate coil state. Disordered regions are indicated by question marks (?). Prediction confidence levels are colour-coded with red representing high confidence and blue low confidence.

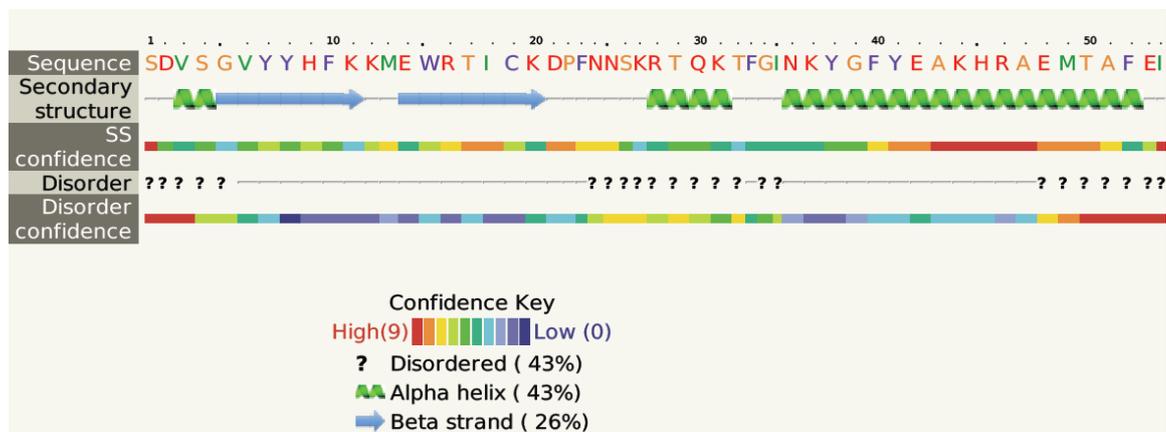


Figure 4.12. Secondary structure prediction of the TA12015 ApiAP2 domain. Blue arrows represent β -strands, green helices indicate α -helices and lines indicate coil state. Disordered regions are indicated by question marks (?). Prediction confidence levels are colour-coded with red representing high confidence and blue low confidence.

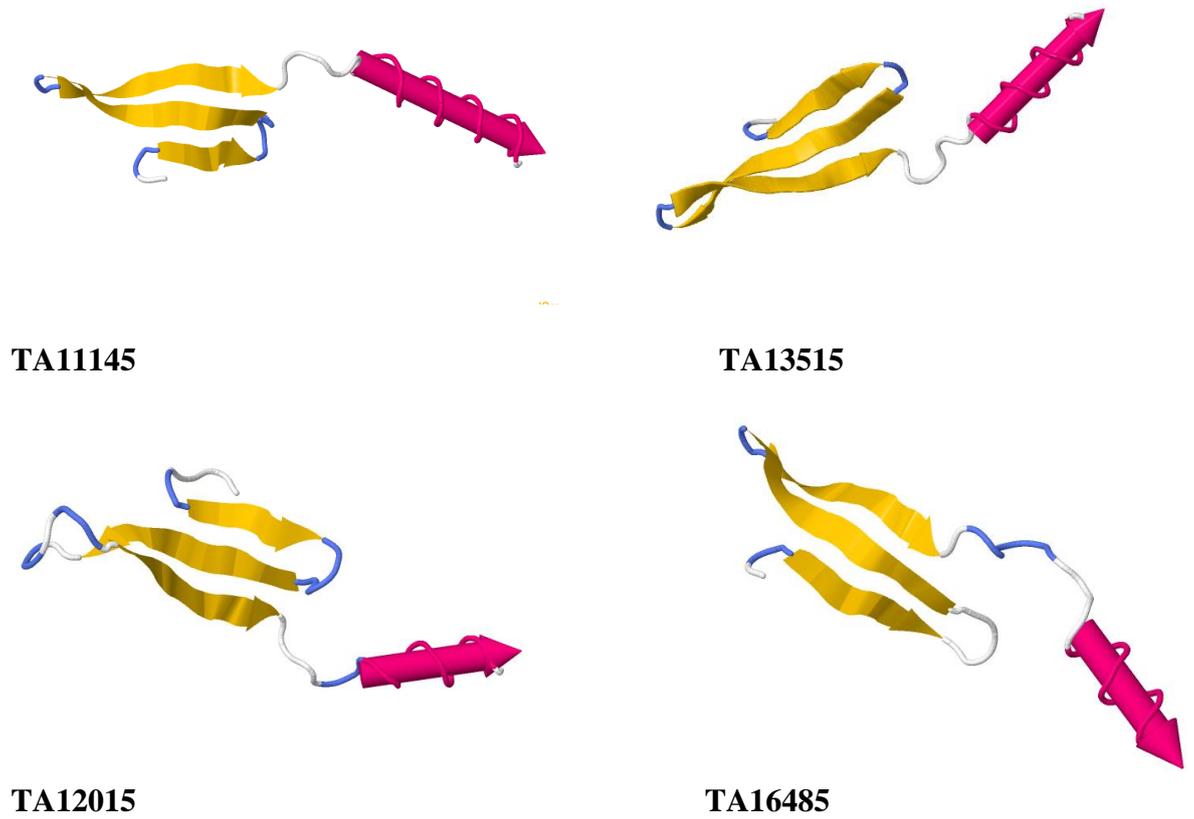


Figure 4.13. A 3-Dimensional structural prediction of four *T. annulata* ApiAP2 domains. Yellow arrows represent β -stands; pink rockets indicate α -helices and lines indicate coil state. Arrowheads point towards the carboxy termini.

4.3.3 Validation of GST-TA13515D AP2 domain binding to its putative binding motif by EMSA

Four GST-fusion proteins (TA13515D, TA11145D, TA16485D and TA12015D) were generated, with TA13515D reaching the highest expression level (Figure 4.14). Generated fusion protein sizes are presented in Table 4.5.

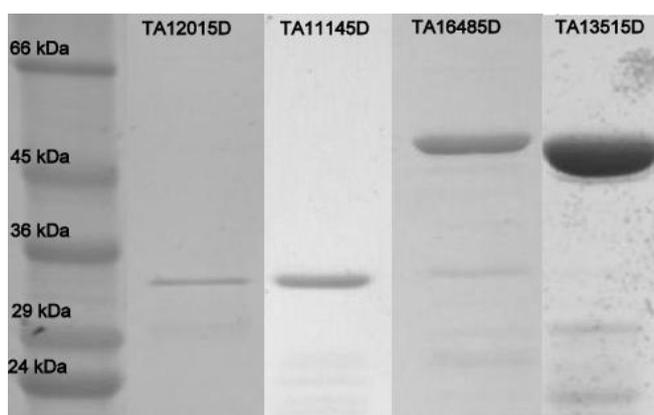


Figure 4.14. GST-ApiAP2 domain fusion proteins generated in this study. GST-TA12015D -34.1kDa, GST-TA11145D - 33.7kDa, GST-TA16485 - 47.6kDa and GST-TA13515D - 45kDa.

ApiAp2 domain	Putative target motif	Domain with extension (kDa)	GST tag (kDa)	GST-DBD (kDa)
TA13515	GTGTAC	19	26	45
TA11145	CACACA	7.7	26	33.7
TA12015	G-box/C-box	8.1	26	34.1
TA16485	TCTACA	21.6	26	47.6

Table.4.5. Properties of the recombinant fusion proteins produced in this study.

Based on ApiAP2 gene orthology and motif enrichment results for gene expression data sets for *T. annulata* presented above, combined with analysis of published data on *Plasmodium* ApiAP2 cis-acting elements, the two ApiAP2 encoding genes of *T. annulata* that showed greatest potential for further study was considered to be TA13515 and TA11145. This was because of the strong identity across the AP2 domain of TA13515 and the *Plasmodium* orthologues together with a data indicating a role in commitment to gametocytogenesis in *Plasmodium*. For TA11145, further interest was based on the early expression profile during merogony and the identification of a phylogenetically related

cluster of AP2 domains that detect CACACA type motifs in *Plasmodium falciparum* (see below).

To validate binding of TaApiAP2 domains to putative motifs recognised by the *Plasmodium* orthologues, EMSA was performed using GST-AP2 domain fusion proteins and biotin labelled double stranded oligonucleotide probes. For the domain of TA13515, EMSA assay was performed with 0.7µg of purified GST-TA13515D and 20fmol of biotin-labelled probe containing the GTGTACAC sequence motif (Figure 4.15). The probe, oligonucleotides kindly donated by Abhinav Sinha (University of Glasgow), was designed based on the motif reported by Campbell et al., (2010). As shown in the Figure 4.15 the recombinant ApiAP2 domain of TA13515 strongly bound to the probe representing the GTGTAC motif with a strong shift of the probe observed (lane 2). No binding was observed with GST alone as a negative control (lane 3). Furthermore the shift was gradually competed with unlabelled probe, with complete competition at 8pmol (Figure 4.16., lane 3-5). No shift was observed with a mutated biotinylated oligo (G/C replaced with A in motif) (lane 6) and with a no motif biotinylated probe (AAAAAAAAA instead of the motif) (lane 7). It can be concluded that the AP2 domain of TA13515 specifically binds to the GTGTAC motif predicted by the *Plasmodium* data.

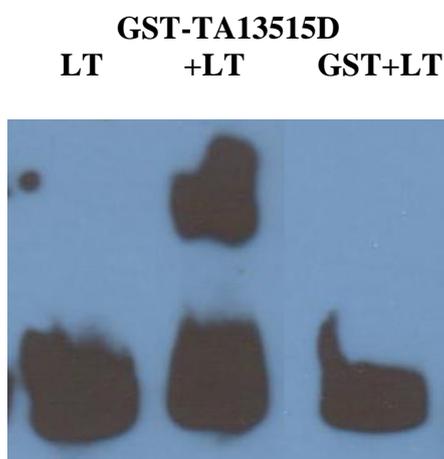


Figure 4.15. EMSA assay performed with 0.7µg of purified GST-TA13515D and 20fmol of a 62bp 5`-biotin-labelled probe containing the GTGTACAC sequence motif. Lane 1 - biotin Labelled Target (LT) probe alone. Lane 2 - GST-TA13515D with Labelled Target. Lane 3 - GST with Labelled Target.

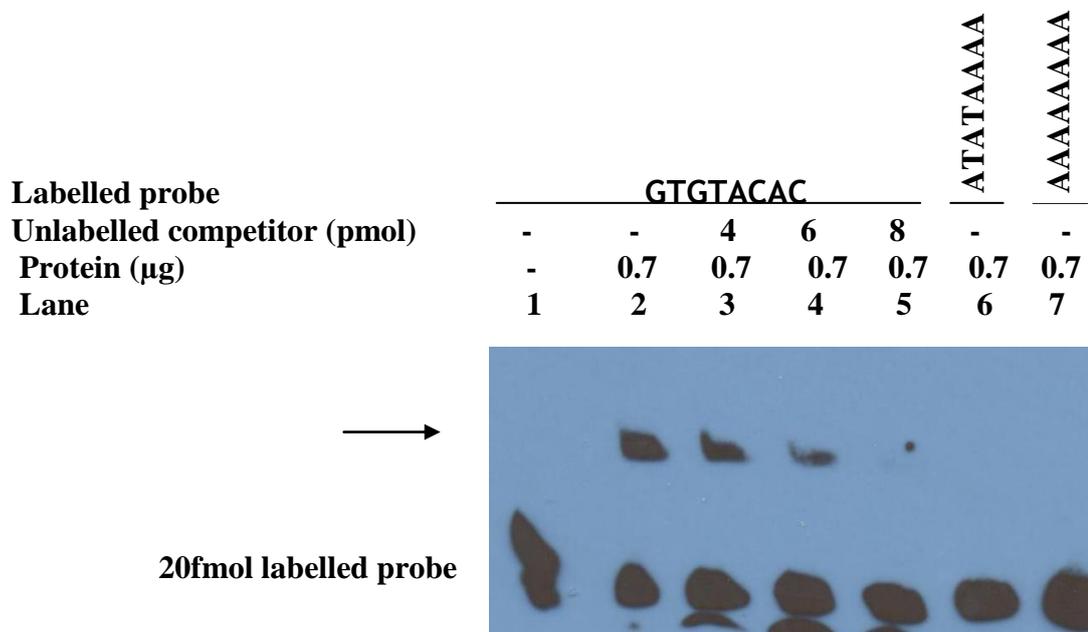


Figure 4.16. EMSA assay performed with 0.7µg of purified GST-TA13515D and 20fmol of biotin-labelled probe containing GTGTACAC motif; specific competitor and mutated probes. Lane 1 - labelled oligo only (GTGTACAC motif). Lane 2 - labelled oligo + GST-TA13515D. Lane 3 - labelled oligo + GST-TA13515D + unlabelled competitor (4pmol). Lane 4 - labelled oligo + GST-TA13515D + unlabelled competitor (6pmol). Lane 5 - labelled oligo + GST-TA13515D + unlabelled competitor (8pmol). Lane 6 - labelled mutated probe (G/C in core sequence replaced with A) + GST-TA13515D. Lane 7 - labelled mutated oligo (no motif - AAAAAAAA instead) + GST-TA13515D. The position of the shift is designated by the arrow.

A Mobility shift assay was repeated for GST-TA13515D protein with the biotinylated DNA sequence obtained from the upstream region of the TA10735 gene (GATA TF) that contained the GTGTAC motif. A specific band shift was observed for GST-TA13515D (Figure 4.17, lane 2) and with the *P. falciparum* AP2-G protein (lane 6). A competition experiment (lane 3-5) showed that unlabelled competitor reduced intensity of the shifted bands; however in this case competition was not 100% .

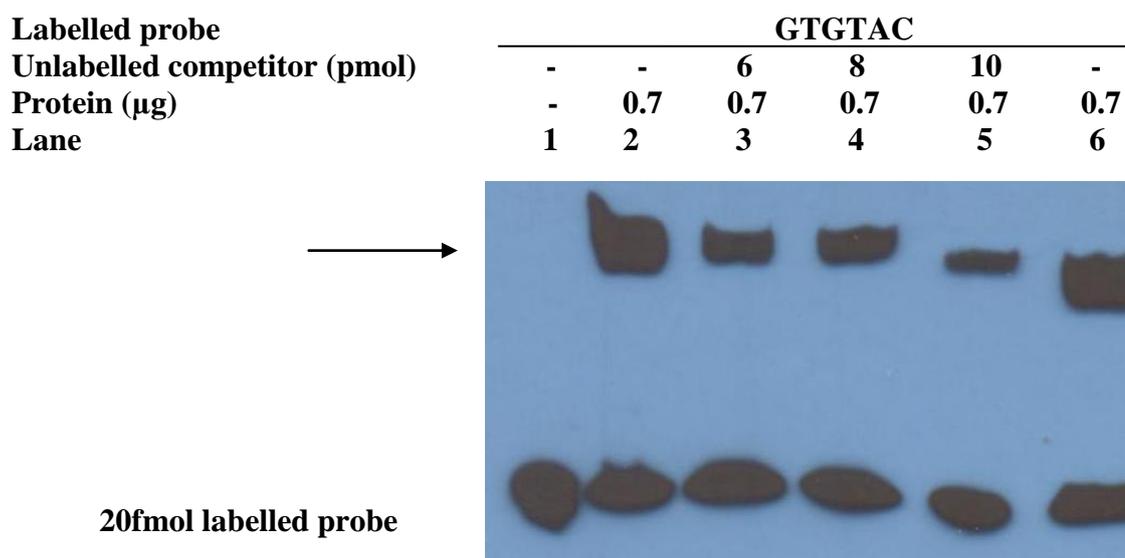


Figure 4.17. EMSA assay performed with 0.7µg of purified GST-TA13515D and 20fmol of biotin-labelled oligo probe containing the GTGTAC motif derived from the upstream region of the TA10735 gene and cold competitor probe. Lane 1 - Labelled GTGTAC oligo. Lane 2 - Labelled oligo + GST-TA13515D. Lane 3 - Labelled oligo + GST-TA13515D + 6 pmol of unlabelled competitor. Lane 4 - Labelled oligo + GST-TA13515D + 8 pmol of unlabelled competitor. Lane 5 - Labelled oligo + GST-TA13515D + 10pmol of unlabelled competitor. Lane 6 - Labelled oligo + PfAP2-G. Position of the shift is designated by the arrow.

Binding to the biotinylated probe containing the GTGTAC motif (putative target motif of TA13515) was confirmed to be protein specific, the band only being obtained when GST-TA13515D protein was added. No interaction was observed between the probe and other GST-AP2 domain fusion proteins. Thus, GST-TA11145D, GST-TA12015, GST-TA16485 proteins did not shift the GTGTAC motif probe (Figure 4.18. lane 3-5). In addition no shift was observed for pure GST protein added to the assay as a negative control (lane 6)

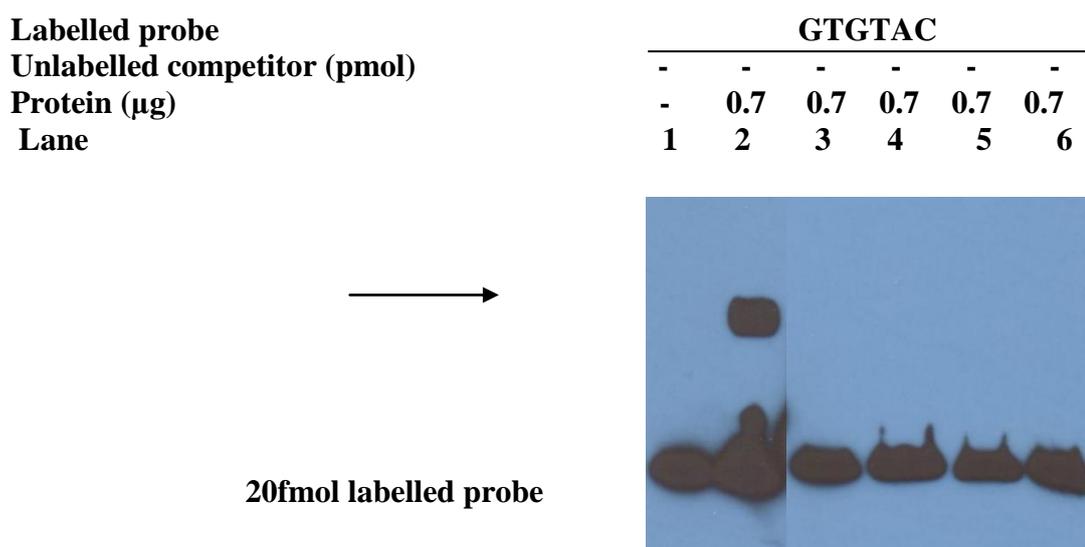


Figure 4.18. EMSA assay performed with 0.7µg of purified GST-AP2 domain fusion proteins and 20fmol of biotin-labelled oligo containing GTGTAC motif derived from the upstream region of the TA10735 gene. Lane 1- labelled oligo only; Lane 2- Labelled oligo probe + GST-TA13515D; Lane 3- labelled oligo probe + GST-TA11145D; Lane 4 - labelled oligo probe + GST-TA12015D; Lane 5 - Labelled oligo probe + GST-TA11145D; Lane 6 - Labelled oligo probe + GST (control). The position of the shift is designated by the arrow.

4.3.3.1 EMSA assay of GST-TA13515D with PA inhibitors

To determine the potential for a synthetic polyamide compound to block ApiAP2 function by competing out the ability to bind to its target motif, EMSA analysis was performed. This was conducted by pre-incubation of the GTGTAC target probe with PA inhibitor, prior to the addition of the TA13515D fusion protein, with a PA designed to bind GTGTAC in the narrow minor groove of DNA (PfAP2-G motif inhibitor ISS-15) or a PA inhibitor designed to recognise the motif TAGCTA

bound by a second *Plasmodium* AP2 domain (PfAP2-O, motif inhibitor ISS-33). Both inhibitors were synthesised in the laboratory of Dr. Glenn Burley and kindly donated by Prof. A. Waters (Glasgow University). The results indicated that the ISS-15 inhibitor specific to GTGTAC prevented the AP2 domain of GST-TA13515D from binding to the biotinylated oligo probe as only faint shifted bands were detected (Figure 4.19, lane 3-4). In contrast, addition of the ISS-33 PA against the unrelated motif, included as a negative control, did not cause a loss in binding between the GST-TA13515D AP2 fusion protein and the target probe. It can be concluded that the ISS-15 PA can specifically compete out binding of ApiAP2 domains to target motifs, as ISS-33 does not recognise the GTGTAC motif and hence no competition is present (Figure 4.20, lane 4-5).

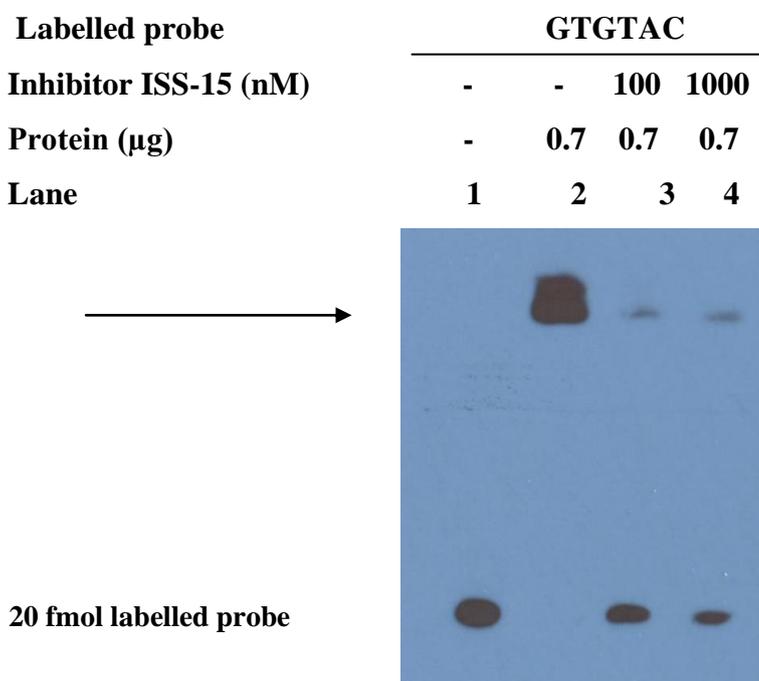


Figure 4.19. EMSA assay performed with 0.7µg of purified GST-TA13515D fusion protein and 20 fmol of biotin-labelled oligo containing the GTGTAC motif derived from the upstream region of TA10735 gene plus the addition of a PA sequence specific inhibitor (ISS-15). Lane 1 - Labelled oligo only; Lane 2 - Labelled oligo + GST-TA13515D; Lane 3- Labelled oligo + GST-TA13515D + ISS-15 (100nM); Lane 4 - Labelled oligo + GST-TA13515D + ISS-15 (1000nM). The position of the shift is designated by the arrow.

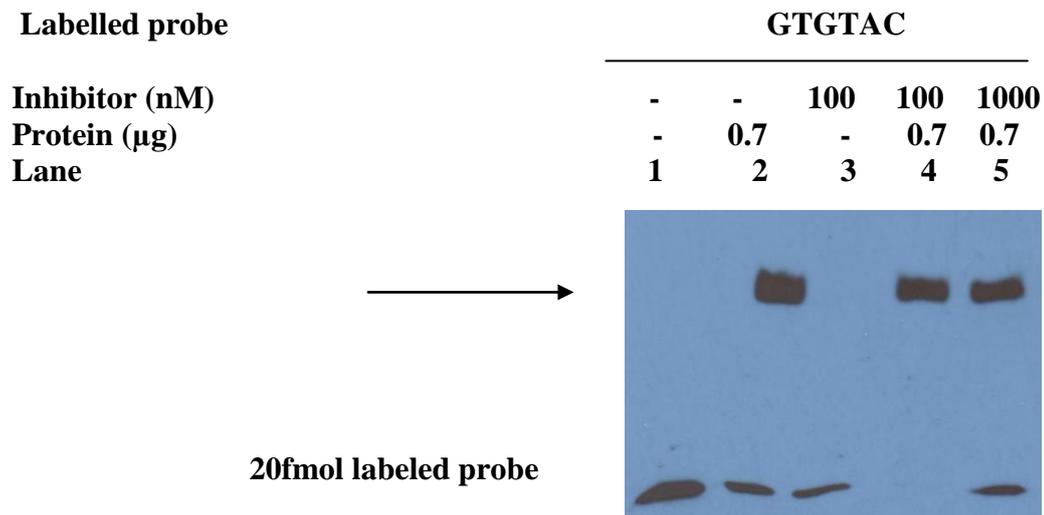


Figure 4.20. EMSA assay performed with 0.7µg of purified GST-TA13515D fusion protein and 20 fmol of biotin-labelled oligo probe containing the GTGTAC motif derived from upstream region of TA10735 gene with the addition of a control PA for non- specific binding (ISS-33). Lane 1 - labelled oligo; Lane 2 - labelled oligo + GST-TA13515D; Lane 3 - labelled oligo + ISS-33 without fusion protein; Lane 4 - labelled oligo + GST-TA13515D + ISS-33 (100nM); lane 5 - labelled oligo + GST-TA13515D + ISS-33 (1000nM). The position of the shift is denoted by the arrow.

4.3.4 EMSA results of the GST-TA11145D fusion protein and its putative binding motif

An initial EMSA experiment was performed with GST-TA11145D and a biotinylated double stranded oligo probe containing a single CACACAC motif. The result indicated binding to the probe but a diffuse shift rather than a tight band was obtained (Figure 4.21, lane 2). No shift was observed when fusion proteins representing unrelated ApiAP2 domains, GST-TA13515D (lane 3), GST-TA12015D (lane 4) and GST-TA16485D (lane 5), were incubated with the probe. Addition of a biotinylated oligo probe lacking a specific AP2 domain motif (AAAAAAAA instead of the CACAC motif) (lane 6), a mutated ATATAAAA motif (lane 7) and an unrelated motif recognized by a distinct ApiAP2 domain, PF14_0633 (TGCATGCA) (lane 8) also failed to show clear specific gel shifts with the CACACAC probe. Thus, although the shift was not tight, it did demonstrate that the GST-TA11145D protein showed specificity for the predicted CACACAC motif. It has been found that specificity of AP2 domain binding is improved using probes derived from upstream region of target genes containing the core motif and

flanking nucleotides. Moreover these types of experiments are relevant to show that the domain could potentially bind to target genes, including the upstream region of the gene encoding the AP2 domain if auto-regulation is implicated. As indicated in Chapter 3 there are multiple ACACAC and CACACA type motifs in the upstream region of the TA11145 that show conservation across *T. annulata* and *T. parva*, implying possible auto-regulation. To determine if TA11145D can bind to these motifs, a probe representing two motifs ACACAC and CACACA separated by five nucleotides and present in IGR of TA11145 was generated (see Appendix - 1.9) and used in EMSA. This probe was called the 2x(A)CACAC(A) probe. The fusion protein generated a strong shift for this probe (Figure 4.22).

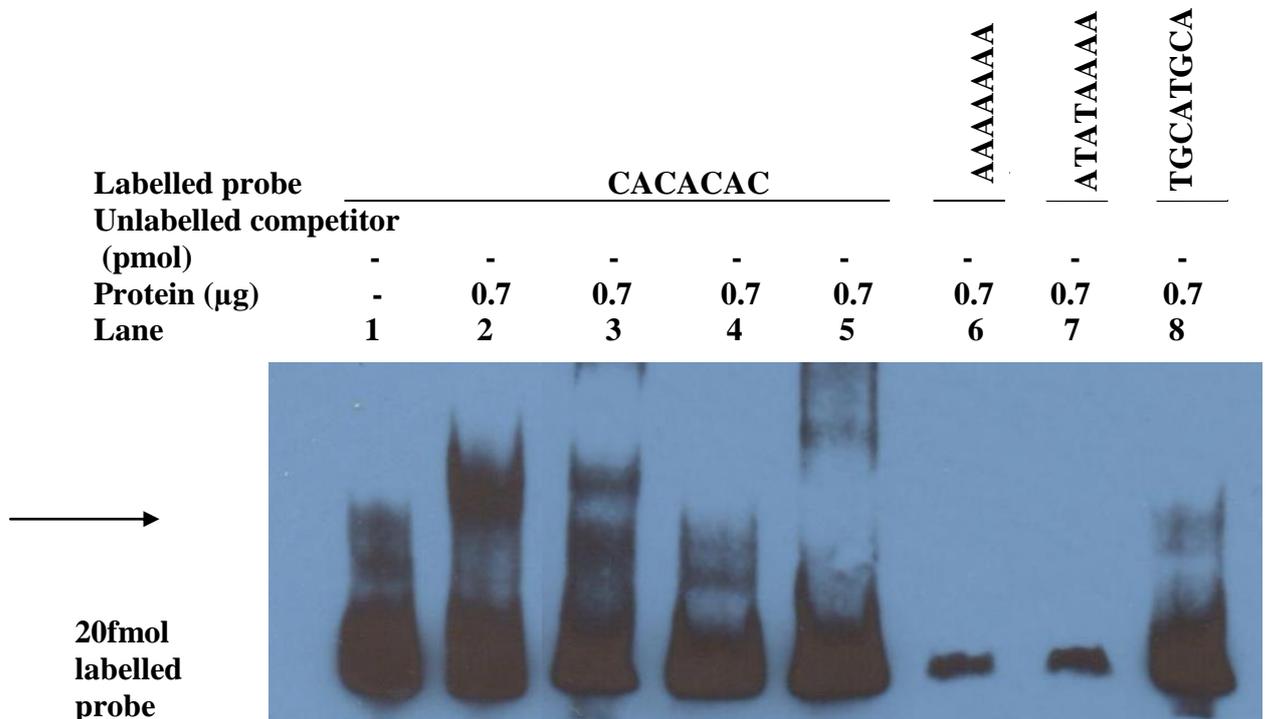


Figure 4.21. EMSA assay performed with 0.7µg of purified GST-ApiAP2 fusion proteins and 20 fmol of biotin-labelled oligo containing the CACAC motif derived from *Plasmodium* AP2 domain binding data. Lane 1, - labelled oligo only; Lane 2, - labelled oligo + GST-TA11145D; Lane 3, -labelled oligo + GST-TA13525D; Lane 4, - Labelled oligo + GST-TA12015D; Lane 5, - Labelled oligo + GST-TA16485; Lane 6, - GST-TA11145D + mutated oligo (AAAAAAA in the place of the CACAC motif); Lane 7, - GST-TA11145D + probe with the ATATAAAA probe; Lane 8, - GST-TA11145D + oligo probe with the TGCATGCA motif. The position of the specific shift is designated by the arrow.

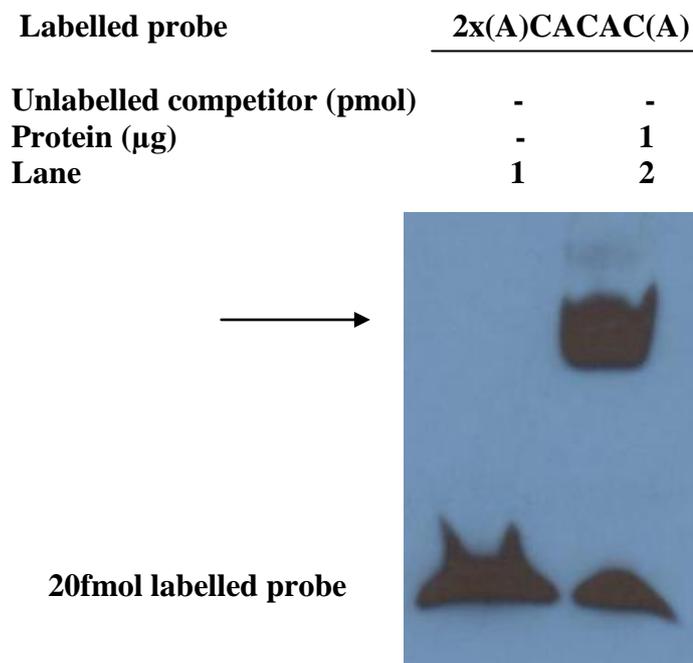


Figure 4.22. EMSA assay performed with GST-TA11145D and labelled oligo probe containing 2x(A)CACAC(A) motifs derived from upstream region of the TA11145 ApiAP2 encoding gene. Lane 1 - Labelled oligo only; Lane 2 - Labelled oligo + GST-TA11145D. Shift position is designated by the arrow.

4.3.5 EMSA results of GST-TA12015 and TA16485D fusion proteins and their putative binding motif

The orthologue of the TA12015 ApiAP2 domain in *C. parvum* (cgd8_810) was found to specifically bind to a G-box motif (Oberstaller et al., 2013). Previously identified by Guo and Silva (2008) and confirmed by MEME analysis (see Chapter 3) TCCCCAT motif is enriched in the upstream regions of *T. annulata* genes implicated in modulation of host cell gene expression that are down-regulated during merogony. Clearly on the opposite strand, the inverse sequence represents a G-box motif. In addition, based on the *C. parvum* data it could be predicted that the AP2 domain of the TA12015 encoded protein could bind a G-box rich motif. TA12015 was difficult to express in a soluble state (Figure 4.23), however and with increasing concentration of Gly-Gly, the amount of soluble GST-TA12015D increased with 500mM of Gly-Gly being the most effective (Figure 4.24).

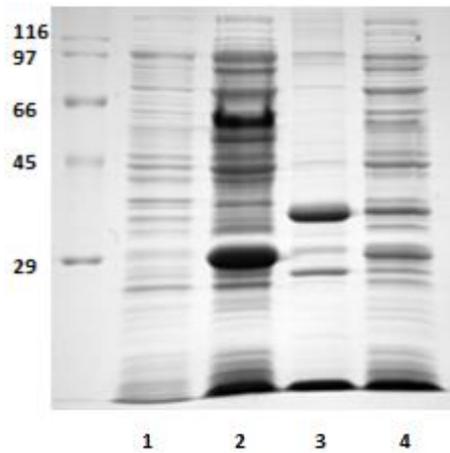


Figure 4.23. GST-TA12015D expression in BL21 cells in TB medium. Lane 1 represents a control (uninduced competent cells in TB medium only). Lane 2 represents soluble GST of 26kDa running at 29kDa, Lane 3 - induced with IPTG insoluble fraction (pellet) with GST-TA12015D fusion protein of 34.1kDa and Lane 4 - induced with IPTG soluble fraction (supernatant) with small amount of GST-TA12015.

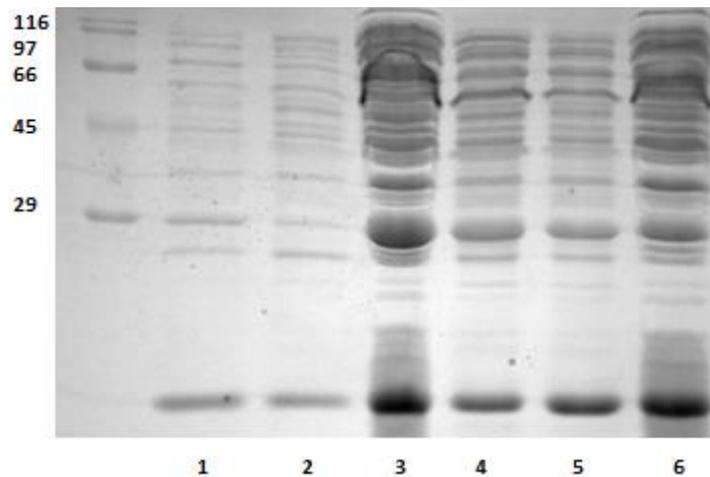


Figure 4.24. Enhanced solubility of the GST-TA12015D in supernatant fraction by induction of protein expression in the presence of the different concentrations of Gly-Gly. Lane 1 represents uninduced cells in TB medium, Lane 2 - supernatant in TB medium, Lane 3 - GST-TA12015D expressed in the presence of 500mM of Gly-Gly, Lane 4 - 200mM Gly-Gly, Lane 5 - 50mM Gly-Gly, Lane 6 - 1M Gly-Gly.

The biotinylated oligo probe represented a TCCCCAT/AGGGGTA motif present in the upstream region of the Tash1-like gene (TA03125) was used in the EMSA assay with GST-TA12015D. However no specific DNA - protein binding was observed (Figure 4.25).

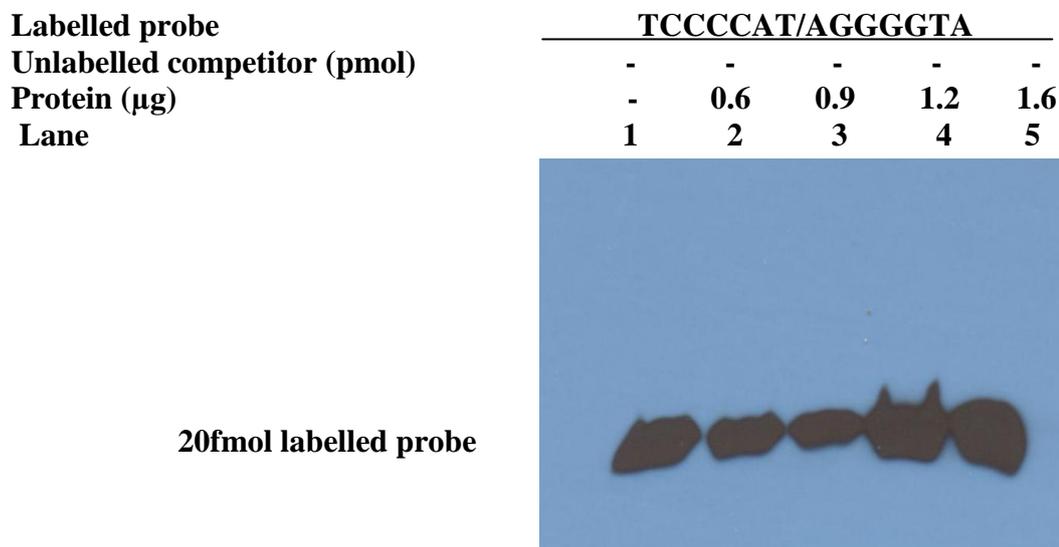


Figure 4.25. EMSA assay performed with labelled oligo probe containing C/G-box motif derived from the upstream region of the Tash1-like gene (TA03125) and increasing concentrations of GST-TA12015D. Lane 1, -labelled oligo only; Lane 2, -labelled oligo + GST-TA12015D (0.6µg); Lane 3, -labelled oligo + GST-TA12015D (0.9µg); Lane 4, -Labelled oligo + GST-TA12015D (1.2µg); Lane 5, -Labelled oligo + GST-TA12015D (1.6µg). No shift was observed.

Gel shift analysis also performed with GST-TA16485D using a biotinylated oligo probe derived from the upstream region of Tash1-like gene, TA03125 containing the putative target DNA core motif identified for the orthologue of this AP2 domain by binding studies performed for *Plasmodium* (Campbell et al., 2010). TCTATA motif is located down-stream of the C-box motif separated by six nucleotides. The EMSA generated a clear band shift (Figure 4.26, lane 2). A competition experiment (Figure 4.38, lane 3-4) showed that unlabelled competitor reduced intensity of the detected shifted bands, however competition was not 100%.

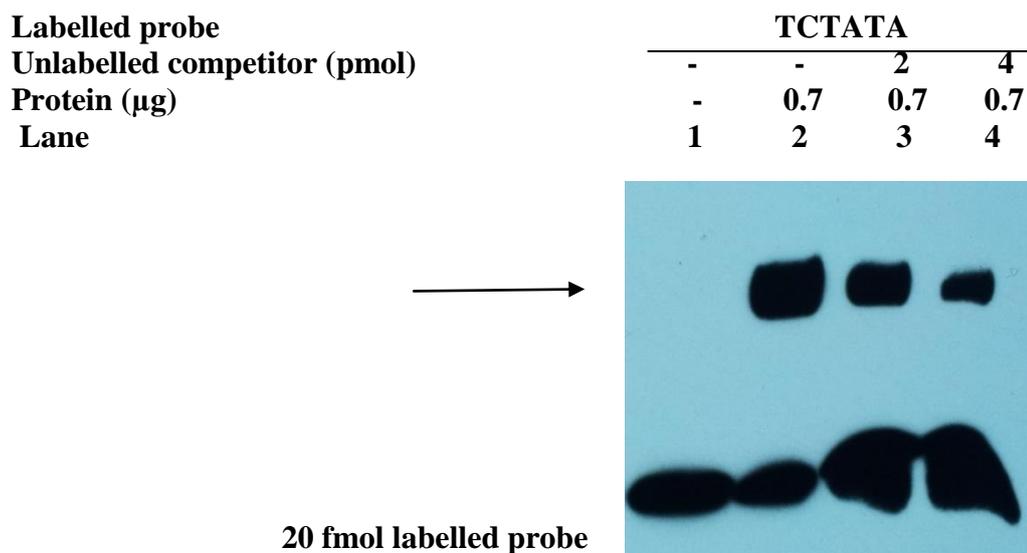


Figure 4.26. EMSA assay performed with GST-TA16485D and biotinylated oligo probe containing the TCTATA motif. Lane 1 - labelled oligo only; Lane 2 - labelled oligo + GST-TA16485; Lane 3 - labelled oligo + GST-TA16485D + cold oligo (2pmol); Lane 4 - labelled oligo + GST-TA16485 + cold oligo (4pmol). Shift position is designated by the arrow.

4.3.6 Detection by EMSA of down-regulated *T. annulata* parasite nuclear factors that bind to the C-box oligonucleotide motif probe

Using MEME and enrichment analysis a C-box motif was identified in the IGRs of TashAT genes family (see chapter 3). The majority of TashAT genes have been shown to be down-regulated during differentiation by Northern blot, microarray or Western blot analysis. However, no *T. annulata* AP2 polypeptide domain could be shown to bind the motif, even though TA12015 was considered as a suitable candidate. To test if a binding activity could be detected for this motif in parasite nuclear extracts, a macroschizont infected cell derived PNE (Day 0) was tested with a probe containing the C-box oligo present in the intergenic regions of the TashAT gene family (sequence from TA03125). As shown in Fig 4.27, a strong probe shift was obtained with PNE from Day 0 culture. Specificity of the band shift was observed when cold competitor was added to the reaction mix (Figure 4.27, lane 3) as cold probe removed the shift while a cold version of the 2x(A)CACAC(A) motif did not compete out the band shift (Figure 4.27, lane 4). It was concluded that G-box/C-box motif is an active cis-binding site for a nuclear factor in *T. annulata*. EMSA was additionally performed with C-box oligo and parasite nuclear

extracts derived from macroschizont stage (Day 0) and differentiating culture (Day 7 and Day 9) (Figure 4.28). Strong down-regulation of the shift (Figure 4.28, line 3-4) was observed in nuclear extracts derived from Day 7 and 9 of differentiating merozoite stage cultures.

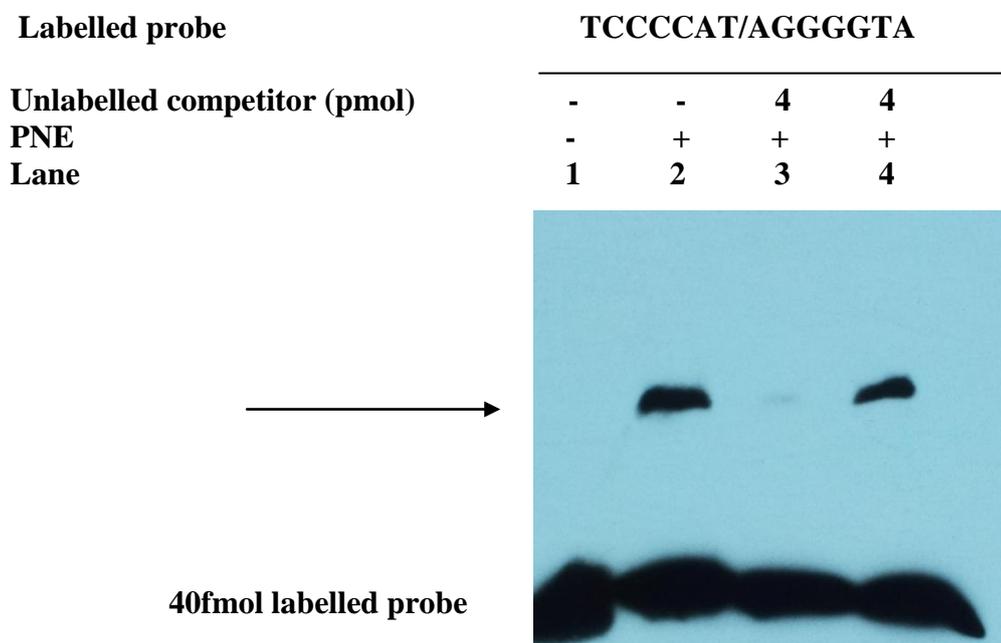


Figure 4.27. EMSA experiment with PNE (D7, Day 0) and the labelled C-box oligo probe derived from the intergenic region of Tash1-like (TA03125) gene with and without cold probe competition. Lane 1 - labelled oligo only. Lane 2 - labelled oligo + D7 PNE Day 0; Lane 3 -labelled oligo + D7 PNE Day 0 + cold C-box probe; Lane 4 - labelled oligo + D7 PNE Day 0 + cold 2x(A)CACAC(A). Shift position is designated by the arrow.

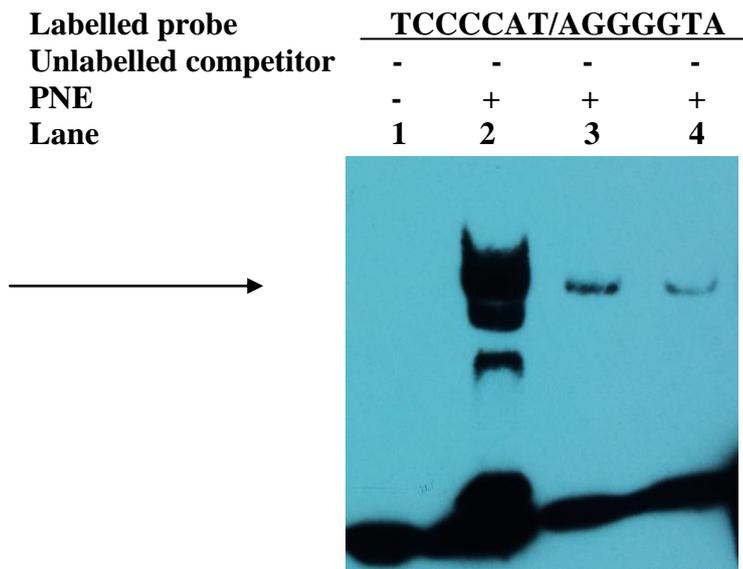


Figure 4.28. EMSA experiment with PNE and the labelled C-box oligo probe derived from the intergenic region of Tash1-like (TA03125) gene. Lane 1 - labelled oligo only; Lane 2 - Labelled oligo + D7 PNE Day 0; Lane 3 - Labelled oligo + D7 PNE Day 7; Lane 4 - D7 PNE Day 9. Shift position is designated by the arrow.

4.3.7 Multiple AP2 domains are predicted to bind (A)CACAC(A) type motifs

Protein binding microarray experiments performed for *Plasmodium* indicated that multiple AP2 domains bind to related versions of the (A)CACAC(A) nucleotide motif (Campbell et al., 2010). Given the identification of a clear orthologue of the up-regulated TA11145 AP2 domain in *Plasmodium* and evidence for enrichment/depletion of (A)CACAC(A) type motifs in the intergenic region of differentially expressed genes in the *T. annulata* data sets, analysis of all potential (A)CACAC(A) type binding domains in the *Theileria* genome was performed. Five *T. annulata* ApiAP2 genes with potential to bind to this motif based on the identification of a *Plasmodium* orthologue in the analysis of Balaji et al. (2005) were found (TA11145 - orthologue of MAL8P1.153 that binds to CACACACA; TA07550, closest to PFD0985w_D2 that binds to GTGTTACAC; TA07100 orthologue of PFD0985w_D1 that binds to ACACAC; TA19920 - orthologue of PF14_0533 that binds to CACACA and TA02615 - orthologue PF13_0026 that binds TGCACACA). Of these genes, based on the microarray

data, only TA11145 was significantly up-regulated during macroschizont to merozoite stage differentiation (Fig 4.29). TA07550 was generally highly expressed during differentiation and showed a small transient increase in mRNA level early in merogony at Day 4. TA07100, TA19920 and TA02615 expression was almost constitutive with only a slight up-regulation of TA19920 from the macroschizont (Day 0) to Day 4, while TA07100 showed down-regulation from Day 7/Day 9 to the piroplasm stage.

Analysis of ClustalW2 sequence alignment of TA11145, TA07550, TA07100, TA19920 and TA02515 ApiAP2 domains with their orthologues from *T. parva* and *T. orientalis* showed high level of conservation of particular residues between species; however differences between particular domains are visible (Figure 4.30). TA02615 ApiAP2 domain and its orthologues showed higher divergence at the amino acid level in comparison to the four other TaApiAP2s. Additionally, the neighbor joining phylogenetic tree with PID (% identity) of the predicted three *Theileria* species and *P. falciparum* was performed (Figure 4.31) with the use of Jalview software. Phylogenetic analysis showed that four putative (A)CACAC(A)-binding *T. annulata* ApiAP2 domains form a separate group distant from the TA02615 domain that is identical to its *P. falciparum* orthologue. This suggests that it may be playing role in later developmental stages, as in *P. berghei*, the (A)CACAC(A)-binding motif for this protein was identified in zygote stage (Campbell et al., 2010). The other motifs TA11145 and TA07100 have clear orthologues in *Plasmodium* and show close relationship to one another. TA07550 and TA19920 do not appear to have a clear orthologue in *Plasmodium* but by blast are most closely related to PF14_0533 which is supporting the tree. Thus, as in *Plasmodium*, there appear multiple phylogenetically related AP2 domains in *T. annulata* predicted to bind (A)CACAC(A) type motifs.

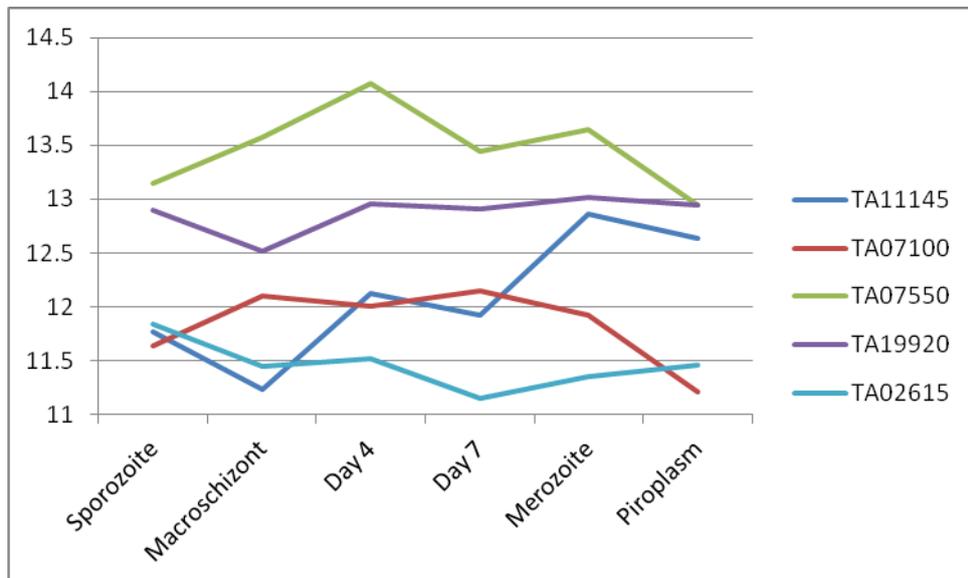


Figure 4.29. Differential expression profiles of five ApiAP2 genes potentially binding CACACA motif in the D7 cell line. Y-axis denotes \log_2 expression values.

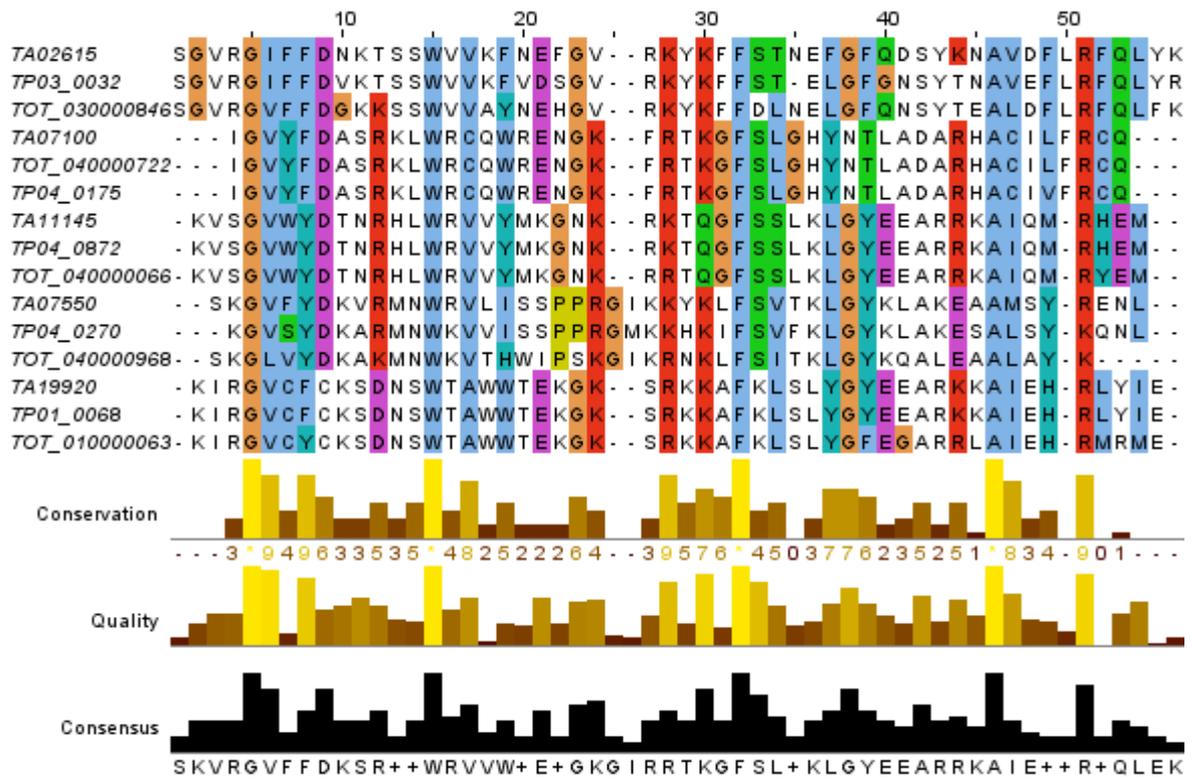


Figure 4.30. *T. annulata*, *T. parva* and *T. orientalis* ApiAP2 alignment of domains that are likely to bind the CACACA motif. Alignment of 5 ApiAP2 factor domains performed using ClustalW showed that these factors demonstrated a high level of similarity between the species and between particular CACACA-binding proteins. Identical residues are highlighted across all the species and proteins. TA02615 and its orthologues showed bigger divergence in on the amino acid level in comparison to four other TaApiAP2s.

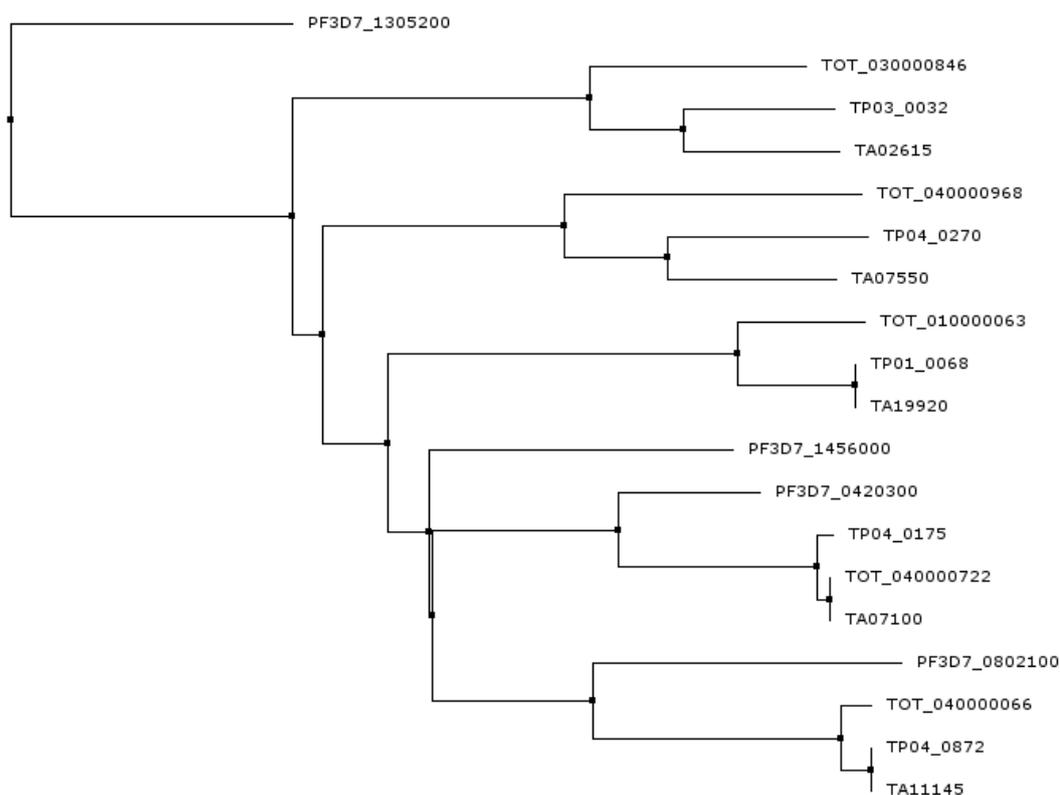


Figure 4.31. Neighbour joining phylogenetic tree of the predicted *Theileria* and *Plasmodium* CACACA-binding ApiAP2 domains. The tree demonstrates that the four TaApiAP2 factors (TA07100, TA19920, TA11145 and TA07550) are more similar to one another than to TA02615 ApiAP2 domain. The tree was created using Jalview, using neighbor joining method with PID (% identity).

4.3.7.1 Detection by EMSA of *T. annulata* parasite nuclear factors that bind to TA11145 oligonucleotide motif probes

In addition to analysis of GST-fusion proteins binding specificity, EMSA was performed with parasite nuclear extract (PNE) generated from the D7 cloned cell lines cultured at 37 °C (Day 0, macroschizont) and following culture at 41 °C (Day 7 and Day 9) to induce merogony. The upstream region of the TA11145 gene containing two (A)CACAC(A) type motifs was used as a probe. This 2x(A)CACAC(A) probe is bound by the TA11145D fusion protein (Figure 4.22). In addition, from analysis of AP2 domain orthologues of *Theileria* and *Plasmodium*, together with confirmation that close orthologues can bind the same motif, this probe is a potential target for another four TaApiAP2s (TA07100, TA07550, TA19920, TA02615). EMSA with the biotinylated 2x(A)CACAC(A) probe and nuclear extracts derived from parasite enriched nuclei at Day 0, Day 7 and Day 9

time points generated four detectable probe shifts (Figure 4.32). The slowest shift (A) was detected with extracts of all three time points, while the next shift complex was clearly detected in extracts of differentiating culture (Day 7 and Day 9) but not in extracts derived from Day 0 (macroscopic infected cells). The two faster shifts C and D were obtained with Day 0 extracts but were clearly either absent or markedly reduced in extracts derived from cultures differentiating to merozoites.

Labelled probe	2x(A)CACAC(A)			
Unlabelled competitor	-	-	-	-
PNE	-	+	+	+
Lane	1	2	3	4

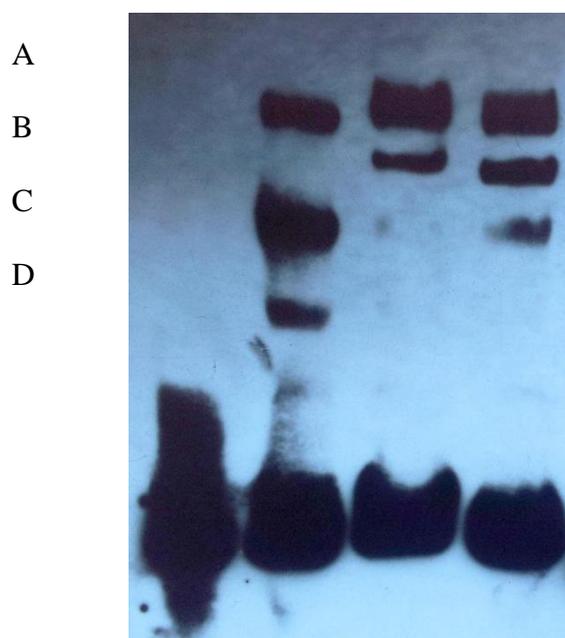


Figure 4.32. EMSA assay performed with D7 Parasite Enriched Nuclear Extract (PNE) from macroschizont (Day 0) and differentiating cultures (Day 7 and Day 9) in the presence of labelled oligo probe containing the ACACAC:CACACA tandem motifs derived from the upstream region of TA11145. Lane 1 - labelled oligo only; Lane 2 - labelled oligo + D7 Day 0 PNE; Lane 3 - labelled oligo + D7 Day 7 PNE; Lane 4 - labelled oligo + D7 Day 9 PNE. A-D - four detected shifts.

EMSA was additionally performed with the D7 parasite nuclear extract (PNE) from Day 0 culture and the TA11145 upstream motif probe to test for specificity by adding unlabelled specific competitor. With probe and extract alone two shifts designated as shift A and C (relative position on repeated EMSA) were obtained (Figure 4.33, lane 2). On addition of unlabelled probe at increasing concentration (2 pmol - lane 3, 4 pmol - lane 4 and 6 pmol - lane 5) reduced

binding at 2 pmol concentration was obtained for shift A. This result was also obtained at 4 and 6 pmol and a reduction of shift C was detected at 4 pmol. Addition of second unlabelled competitor was also performed (TA03125 upstream region probe containing C-box and TCTATA motifs) (lane 7-9). This cold probe showed evidence for a reduction in the level of both A and C shifts compared to the reaction without competitor: it should be noted that the probe contains an ACA motif on its reverse strand and therefore may be expected to show a level of competition relative to the labelled probe (see probe sequence in Appendix - 1.9). It was concluded that shift A showed greater specificity for the TA11145 probe than shift C under these conditions, although greater abundance of shift C relative to A may have influenced the result.

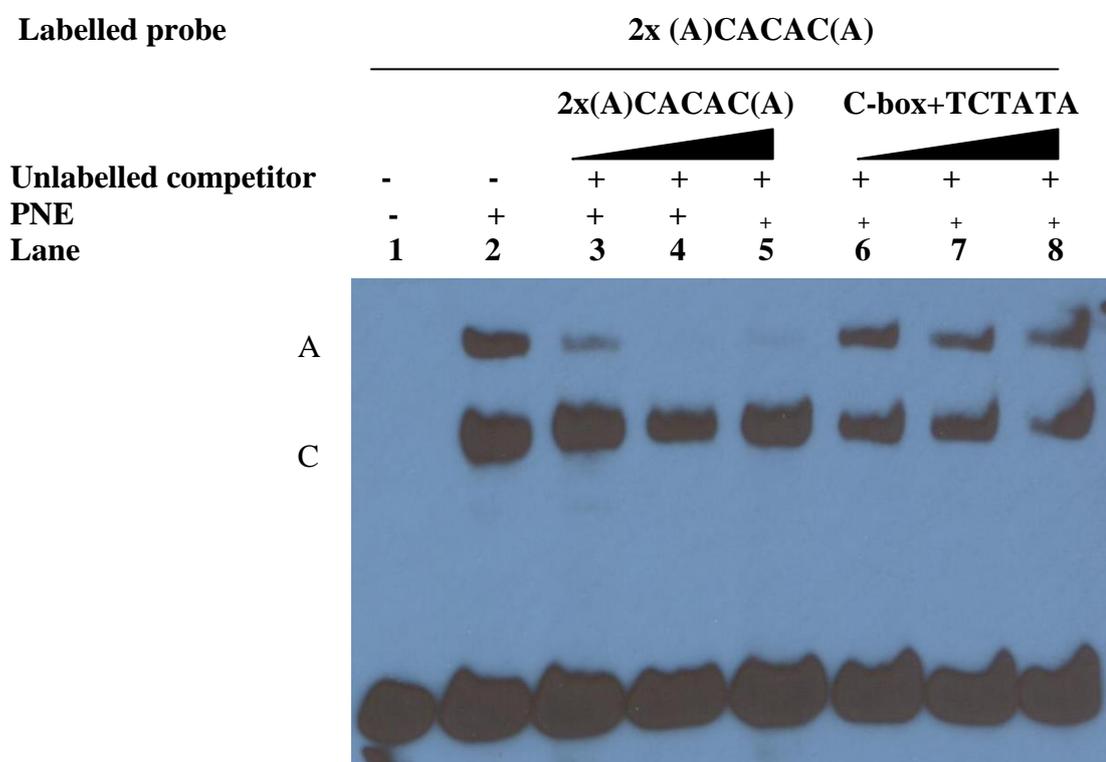


Figure 4.33. EMSA assay performed with D7 Parasite Nuclear Extract (PNE) from macroschizont (Day 0) in the presence of labelled oligo containing two (A)CACAC(A) motifs derived from the upstream region of TA11145 and specific and non-specific competitors. Lane 1 - Labelled oligo only; Lane 2 - Labelled oligo + D7 Day 0 PNE; Lane 3 - Labelled oligo + D7 Day 0 PNE + specific competitor (2pmol); Lane 4 - Labelled oligo + D7 Day 0 PNE + specific competitor (4pmol); Lane 5 - Labelled oligo + D7 Day 0 PNE + specific competitor (6pmol); Lane 6 - Labelled oligo + D7 Day 0 PNE + non-specific competitor (2pmol); Lane 7 - Labelled oligo + D7 Day 0 PNE + non-specific competitor (4pmol); Lane 8 - Labelled oligo + D7 Day 0 PNE + non-specific competitor (6pmol).

To confirm that binding was related to the (A)CACAC(A) motifs in the TA11145 probe, nucleotide substitution in the ACACAC and CACACA motifs was performed to generate the mutant probes shown in Table 4.6.

Oligo name	Oligo sequence
2x(A)CACAC(A) oligo	GAT <u>ACACAC</u> TTATG <u>CACACACA</u>
(A)CACAC(A)-MUT oligo	GAT <u>ACACAC</u> TTATG <u>CAGAATAT</u>
(A)CACAC(A)-2xMUT oligo	GAT <u>ATAGAA</u> TTATG <u>CAGAATAT</u>

Table 4.6. Mutated biotinylated oligos used in EMSA experiment obtained from upstream region of TA11145 gene.

Using the double motif mutated probe, none of the 4 mobility shifts that were obtained with the control wild type probe could be detected (Figure 4.34, lane 7-8). To test if there was any binding to the probe when only one of the motifs was present, a second mutant probe where motif CACACA was altered to CAGAAT was used. In contrast to the double mutant, shift complex B (upper part of doublet with non mutated probe) detected with PNE derived from differentiating culture, was clearly observed with the mutant 2 probe (Figure 4.34, lane 5-6). A markedly reduced level of binding was observed for the other three shifts (absent for A and D), with shift C still detectable at a faint level in Day 0 PNE (lower part of doublet with shift B). These results indicate differential specificity of different complexes to the two motifs in the probe, with the up-regulated shift complex B showing preference for the ACACAC motif, or there is a greater requirement for both motifs by the other 3 complexes.

Labelled probe	ACACAC+CACACACA				ACACAC+ CAGAATAT		ATAGAA+ CAGAATAT	
	-	-	-	-	-	-	-	-
Unlabelled competitor	-	-	-	-	-	-	-	-
PNE	-	+	+	+	+	+	+	+
Lane	1	2	3	4	5	6	7	8

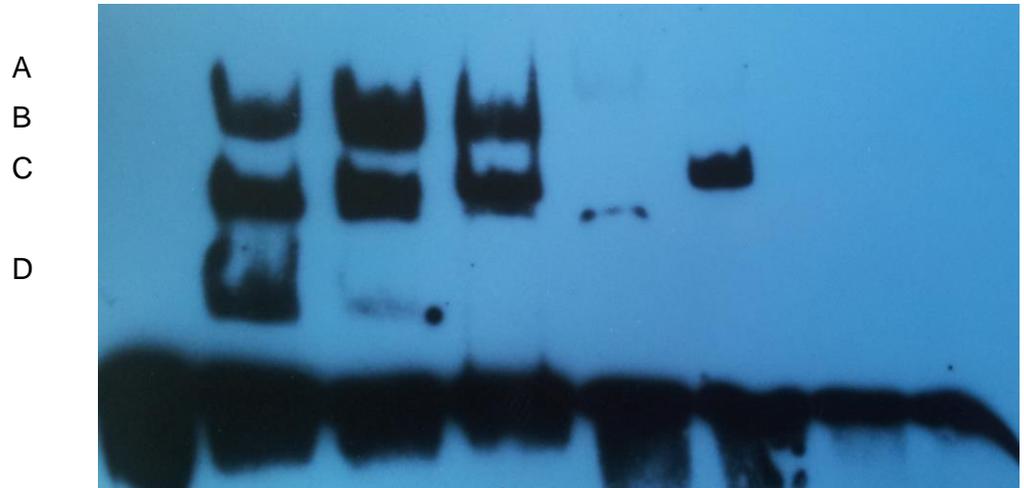


Figure 4.34. EMSA assay performed with D7 Parasite Nuclear Extract (PNE) and TA11145 upstream ACACA-CACACA labelled probe and mutated versions of this probe. Lane 1 - Labelled 2x(A)CACAC(A) oligo only; Lane 2 - Labelled 2x(A)CACAC(A) oligo + PNE Day 0; Lane 3 - Labelled 2x(A)CACAC(A) oligo + PNE Day 7; Lane 4 - Labelled 2x(A)CACAC(A) oligo + PNE Day 9; Lane 5 - Labelled mutant probe where motif CACACA was altered to CAGAAT + PNE Day 0; Lane 6 - Labelled mutant probe where motif CACACA was altered to CAGAAT + PNE Day 9; Lane 7 - Labelled double mutant probe + PNE Day 0; Lane 8 - Labelled double mutant probe + PNE Day 9.

4.3.7.2 Comparative analysis of two putative (A)CACAC(A)-binding ApiAP2 proteins in cell lines competent or attenuated for differentiation to the merozoite

T. annulata-infected clonal cell lines D7 and D7B12 have the same genetic background; however while in the D7 cell line, *T. annulata* differentiates from the macroschizont to the merozoite at 41 °C, the D7B12 cell line is severely attenuated for the differentiation process. Altered gene expression levels at the macroschizont life cycle stage between these two cell lines may reflect an altered baseline state related to their potential to differentiate and could identify possible candidate regulators of the differentiation process: a hypothesis that is tenable as expression of the merozoite surface antigen gene, TamS1 is detectable at this stage and is higher in D7 than D7B12 (Shiels et al., 1994).

To test for differential expression of ApiAP2 genes between D7 and D7B12 analysis of microarray data representing macroschizont infected cells for both lines was performed by Dr W. Weir using a Rank Product approach (Table 4.7). The Rank Product results showed that expression levels of three AP2 domain genes showed a significant reduced expression level (>2 fold difference, FDR<0.5) in the D7B12 line that is attenuated for differentiation. Of interest, two of these ApiAP2 genes TA11145 and TA07550 bind or are predicted to bind (A)CACAC(A) motifs: a 3.95 fold decrease for TA11145 and a 3.23 fold change decrease for TA07550 in the D7B12 line compared to the D7 cell line. In contrast, the three other potentially (A)CACAC(A) binding ApiAP2s (TA07100, TA19920 and TA02615) did not show a significant difference in expression levels between D7 and D7B12 cell lines (1.13 fold reduction in expression in D7B12 for TA07100, TA19920 and 1.51 for TA02615). Based on the above data, two ApiAP2s (TA11145 and TA07100) were chosen for further validation by qRT-PCR to assess differences in gene expression levels between the cell lines. These two genes represent the most up-regulated (TA11145) and the most down-regulated (TA07100) of the five (A)CACAC(A) ApiAP2 encoding genes during differentiation from the macroschizont to piroplasm.

Gene ID	D7 vs D7B12	D7 vs D7B12
	FC	Min FDR
TA07550	-3.23	0.01
TA13395	-1.10	1.20
TA07100	-1.13	1.21
TA19920	-1.13	1.21
TA18095	-1.14	1.21
TA02615	-1.51	0.82
TA09965	-1.55	0.75
TA16535	-1.66	0.49
TA11665	-2.00	0.21
TA11145	-3.95	0.01
TA04435	-7.38	0.00
TA16485	-1.45	0.81
TA17415	-1.56	0.74
TA16105	-2.42	0.06
TA08375	-1.72	0.47
TA12015	-1.79	0.41
TA05055	-1.51	0.79
TA04145	1.00	1.02
TA13515	-1.87	0.32

Table 4.7. Differential expression of TaApiAP2 genes between D7 and D7B12 cell lines (FC - fold change (\log_2), Min FDR - minimum false discovery rate). Highlighted in blue - putative TaApiAP2s binding to (A)CACAC(A) motif.

qRT-PCR results confirmed the TA11145 microarray results showing a change in relative difference in expression level (Figure 4.35) for these two genes across the D7/D7B12 comparison. Thus at the D0 time point there was an approximately 42 fold (absolute) increased expression in TA11145 in the D7 line relative to D7B12 and this difference increased, as might be expected to 427 fold at Day 4 and 2076 fold at Day 7. In contrast, differences in expression level of TA07100 were less marked with a relative increased expression in D7 vs D7B12 of 1.23, 1.55 and 4.92 fold at Day 0, 4 and 7, respectively. In addition analysis of expression levels indicated that TA07100 is expressed at a higher level in D7, Day 0 cells. This difference in favor of TA07100 over TA11145 was found to be, as

expected, increased in the D7B12 line at the Day 0 time point (see Appendix 4.1). This difference was maintained when D7B12 cells were placed at 41 °C but the ratio was significantly altered in favor of TA11145 in the D7 line, when induced to differentiate at the elevated temperature.

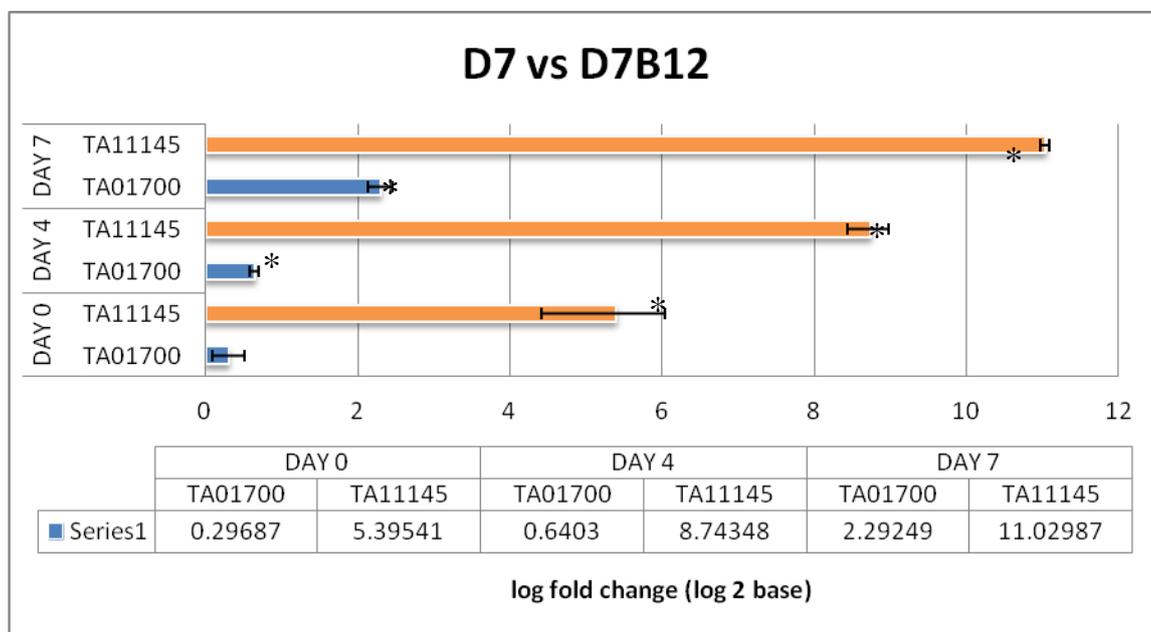


Figure 4.35. Comparison of expression levels of TA11145 and TA07100 ((A)CACAC(A)-binding ApiAP2 genes) of particular time-points between D7 and D7B12 cell lines based on qRT-PCR data. Increased expression of TA11145 (orange colour) in D7 cell line during differentiation event is clearly detectable (5.4 for Day 0; 8.7 for Day 4 and 11.02 for Day 7 fold difference in log₂ scale). Small increase in expression of TA07100 (blue colour) in D7 vs D7B12 cell line is visible (0.29 for Day 0, 0.64 for Day 4 and 2.29 for day 7). Observed differences between cell lines were significantly lower for TA07100 compared to TA11145. The asterisk (*) next to the bars indicates statistical significance (two-tailed Student's t-test, p-value≤0.05) relative to macroschizont stage.

Finally, to test whether there may be a difference in the level of nuclear factors that bind to the (A)CACAC(A) type double motif upstream of TA11145 an EMSA experiment using macroschizont enriched nuclear extracts for D7 and D7B12 was performed. The results revealed presence of bands A and C in PNE (Day 0) derived from the D7 cell line. In contrast with PNE derived from the D7B12 cell line (Day 0) only band C was present and a slight reduction of intensity in comparison to D7 culture was observed for this band (Figure 4.36; lane 3). Band D is not always seen in D7 extracts and it might be due to host origin or run variability. It is equally possible that all four bands are of parasite origin and are binding to the (A)CACAC(A)-type motif, but this will require further

investigation, such as gel shifts with host-nuclear extracts from all analysed *Theileria* life-stages.

Labelled probe	<u>2x(A)CACAC(A)</u>		
Unlabelled competitor	-	-	-
PNE	-	+	+
Lane	1	2	3

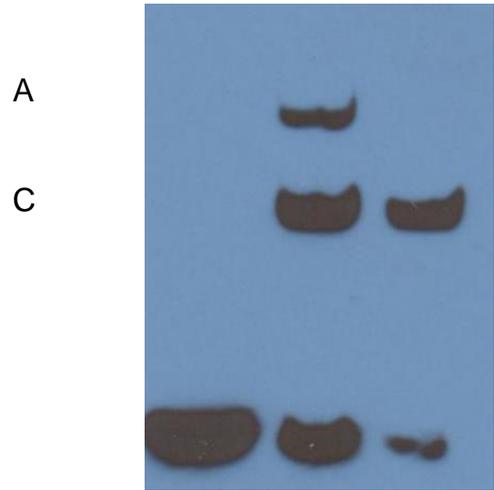


Figure 4.36. EMSA assay performed with D7 and D7B12 Parasite Nuclear Extracts (PNEs) and the TA11145 upstream ACACAC-CACACA labelled probe. Lane 1 - non mutant 2x(A)CACAC(A) probe only; Lane 2 - 2x(A)CACAC(A) probe + D7 Day 0 PNE; Lane 3 - 2x(A)CACAC(A) motif + D7B12 Day 0 PNE.

4.3.8 IFAT results

To detect native AP2 domain proteins in *T. annulata* attempts were made to raise antisera against purified fusion protein representing TA13515D and TA11145D. Antisera were produced in rabbits under commercial contract by Scottish National Blood Transfusion Service. To test for specificity reactivity, Indirect Immunofluorescent Analysis (IFA) was performed with both TA13515D and TA11145D anti-sera. With sera against TA13515, a punctate staining pattern could be detected in slides of the purified piroplasm stage but it was difficult to conclude that the staining co-localised with the nuclei. For sera against TA11145 there was strong detection of parasite nuclei and an elevation of reactivity in parasites differentiating to the merozoite. However, pre-immune serum from one rabbit showed an almost identical pattern and it was impossible to conclude that the observed reactivity was due to detection of the TA11145 encoded protein. Similar results were obtained by Western Blot. It was unknown why the pre-screen IFA failed to show this obvious reactivity with this PI serum, but was most likely due to a technical error. Selected sera were run over a GST column to remove Ab against GST but no improvement was observed in terms of specificity. One possibility for the observed reactivity of the PI serum is that this rabbit had been infected previously with an apicomplexan parasite. Cross recognition of AP2 domains or other conserved nuclear proteins might then be predicted.

4.4 Discussion

Identification of the DNA-binding specificity of ApiAP2 proteins is very important for exploring their role as transcriptional regulators during all stages of parasite development. Differential expression of ApiAP2 encoding genes in specific developmental stages of apicomplexans such as *Plasmodium* and *Toxoplasma* strongly suggests that they may also play an essential role in mediation of transcriptional regulation of *Theileria* stage specific genes. It has been hypothesised that a complex regulatory system operates to regulate differentiation in *Plasmodium* with evidence for epigenetic control, specific TFs and post-transcriptional regulation of gene expression. While it is likely that similar mechanisms also operate in other Apicomplexa species, in this study focus was primarily made on characterisation of selected specific TaApiAP2 transcription factors with validation of specific binding to predicted DNA motifs. The aim of this work was to lay the foundation for further exploration of their role in regulation of stage specific gene expression during merogony of *T. annulata*.

The closest orthologue of TA13515 is PFL1085w (named as AP2-G by Sinha et al., 2014) that binds the motif GxGTACxC (where x denotes any residue) and is now known to be essential for commitment to gametocytogenesis in *Plasmodium* (Sinha et al., 2014; Kafsack et al., 2014). Strong conservation to the *Plasmodium* ApiAP2 domain suggested similar binding specificities for TA13515, while strong up-regulation from macroschizont to merozoite/piroplasm indicated a role in regulation of differential gene expression during or after stage differentiation. Strong binding of GST-TA13515 fusion construct to the GTGTAC motif was confirmed and addition of specific competitor to the EMSA assay completely reduced the binding affinity of this cis-element - protein complex. A binding site upstream of the down-regulated gene encoding a GATA transcription factor was also validated by EMSA for the TA13515 domain. Additionally, EMSA performed with mutated biotinylated oligo (G/C nucleotides were replaced with A) gave no shift, indicating the importance of these nucleotides in specific binding to TA13515D. Finally, a polyamide inhibitor designed to preferentially bind to

GTGTAC motif sequence prevented formation of the DNA-protein complex, providing further evidence of motif specificity for this ApiAP2 protein.

The GTGTAC motif was found in three copies in the upstream region of TA13515, together with one copy of a (A)CACAC(A) motif and 4 copies of a G-box/C-box, suggesting auto-regulation of this ApiAP2, as well as possible regulation by other ApiAP2 factors. It has been postulated that AP2-G expression is auto-regulated (Kafsack et al., 2014) by a positive feedback loop mechanism that commits parasites to production of this ApiAP2 and gametocyte production. As postulated previously (Shiels et al., 1999) this model may also operate in *Theileria* and across the Apicomplexa. This is supported by conservation of the motifs and ApiAP2 domains in factors showing related kinetics of expression during stage differentiation events. Moreover, the presence of many copies of cis-binding elements upstream of TA13515 may indicate that simple binding of one factor to the DNA motif is not sufficient for control of this gene. This emphasises the potential complexity of combinatorial binding of several ApiAP2s and the requirement of additional TFs to regulate parasite gene expression during stage differentiation events.

EMSA experiment performed to detect the presence of G/C-box binding proteins in the parasite enriched nuclear extract revealed a shift specific for the macroschizont stage, indicating that this motif is likely to be an active binding site for *Theileria* nuclear factor(s) down-regulated during differentiation to the merozoite. This motif was previously detected as enriched in the upstream regions of *Theileria* genes (Guo and Silva, 2008) and is significantly enriched in the data set of down-regulated genes encoded by the TashAT cluster. A further variant of this motif was also significantly enriched upstream of SVSP family genes and genes down-regulated during differentiation to the merozoite, but EMSA was not performed with this motif. It is highly likely that the C-box rich motif and unknown nuclear factor(s) operate in regulation of differential expression of the TashAT cluster and other genes down-regulated during merogony. Unfortunately, no specific shift was detected for the GST-TA12015D fusion protein. This may have been due to poor quality (in terms of functionality/domain folding) of the generated fusion protein. It is also possible

that, similar to *P. falciparum*, the presence of only one G-box motif is not sufficient for high affinity binding (Campbell et al., 2010). Finally, a G-box motif may not be a target for TA12015 at all. The possibility that TA12015 represents the native protein that binds to the C-box motif of TashAT cluster genes (inverted G box) on EMSA is unlikely, as this factor appears to be down-regulated whereas the mRNA for TA12015 is up-regulated during merogony. Nevertheless, further study of the binding affinities of this ApiAP2 and its role in merogony are warranted.

ApiAP2 encoding gene TA16485 is an orthologue of PFL1900w that binds to the TCTA[C/T]A motif in *Plasmodium* (Campbell et al., 2010). This domain was found to have the highest level of domain conservation of the studied TaApiAP2s in comparison to its *P. falciparum* orthologue, confirmed by reciprocal BLAST. However, 59% of the whole TA16485 protein structure was predicted as distorted, preventing meaningful characterisation of secondary and tertiary structure. EMSA showed strong binding of the TA16485 ApiAP2 domain to the upstream region of a Tash1-like gene (TA03125) containing a TCTATA and a C-box motif. The addition of cold oligo reduced the binding affinity of the complex. Furthermore, a strong negative Pearson correlation between TashAT/down-regulated from macroschizont to merozoite genes and TA16485 expression (-0.92, see Chapter 3, Figure 3.23B, D) raises the possibility that TA16485 could act to regulate this subsets of genes, with repressor activity being most likely. The network of ApiAP2 genes in the IDC of *P. falciparum* also indicates function as a repressor: as the ApiAP2 domain encoding gene is expressed later in the cycle than its predicted target genes. However, further investigation on TA16485 binding specificity and its target genes is required.

The TA11145 protein is an orthologue of *P. falciparum* MAL8P1.153 ApiAP2 that binds to a (A)CACAC(A)-type motif. Based on orthology and phylogenetic comparison with *Plasmodium*, four other genes in the *T. annulata* genome (TA07550, TA07100, TA19920 and TA02615) can also be considered as putative (A)CACAC(A)-binding factors. Additionally, it is possible that some ApiAP2s have secondary binding motif specificity, as *Toxoplasma* AP2IX-9 (TGME49_306620), orthologue of TA19920 binds to a CAGTGT motif (Radke et al., 2013): alternatively the AP2 domain of *Theileria/Plasmodium* orthologues may have

diverged from *Toxoplasma*. On comparison of expression levels, using microarray data between differentiation competent (D7) and deficient (D7B12) cell lines, TA11145 and TA07550 showed significantly ($FD > 2$, FDR, 0.5) higher relative expression levels in the D7 cell line at Day 0. This result could indicate their importance for macroschizont to merozoite differentiation, together with TA11145 being one of the four most up-regulated ApiAP2 genes during this event.

Further confirmation that expression of TA11145 was elevated in D7 relative to D7B12 was achieved by performing qRT-PCR (Figure 4.35). qRT-PCR was performed relative to TA07100, a predicted (A)CACAC(A) binding ApiAP2 gene that was expected to be constitutively expressed from macroschizont to merozoite and down-regulated at the piroplasm stage. It was clear that TA11145 showed higher expression in D7 than D7B12 cells at the macroschizont stage, while there was no significant difference in TA07100 values, and as expected this higher level of TA11145 expression was exacerbated as cells progressed through Day 4 towards the commitment to differentiate. While a rise in expression was also detected for TA07100 at 41 °C in the D7 line, the differential in expression compared to the D7B12 line was significantly lower than that computed for TA11145 (see Figure 4.35). These results indicate that there is differential expression of AP2 factors that are predicted to bind related ACACAC/CACACA motifs that is associated with the ability of an infected cell line to undergo stage differentiation. Moreover the ability to differentiate is also associated with the change in the ratio of their expression level in favor of TA11145. While qRT-PCR was not performed for the TA07550 gene it is also likely, based on the array data, that it is elevated in the differentiation competent line while TA19920 is less likely to show a difference. Furthermore, gel shift experiment revealed difference in binding potential of nuclear factors to (A)CACAC(A) motifs between D7 and D7B12 cell lines. These results lend evidence to postulation that that these ApiAP2 genes encode candidate regulators of differentiation to the merozoite stage. Similar results of elevated expression of the AP2-G regulator in lines competent for gametocytogenesis

relative to lines deficient for this commitment step was found by Sinha et al., (2014).

The reduced ability of ApiAP2 factors to increase expression level in the attenuated D7B12 line supports the previous stochastic model incorporating an auto-regulatory loop and quantitative commitment point (Shiels et al., 1994). In this model it is more difficult for D7B12 cells to attain this point. Also stochastic activation can be related to a mechanism of flexible baseline protein production, which can be modulated in response to environmental changes (Sinha et al., 2014; Kafsack et al., 2014). The detection of low-level expression of merozoite genes in the preceding macroschizont stage (Shiels et al., 1994, Shiels et al., 2000) together with evidence for elevated expression of ApiAP2 factor genes in the macroschizont stage (one of which is associated with merogony) in the D7 line relative to D7B12 add further support that a stochastic mechanism operates, and that AP2 factors are involved.

The mechanism for altered expression of the ApiAP2 genes between the two cell lines remains unknown. It may well be linked to epigenetic changes causing altered ApiAP2 access to critical binding sites. Activation of gene expression requires remodeling of chromatin structure and opening up the regions that are involved in DNA binding (Constanze and Cockerill, 2000). Modifications of parasite chromatin in the D7B12 cell line, or a lack of them, may block access for binding of critical ApiAP2s to their sites. It is also possible a point mutation in either a regulatory domain, or the motif bound by a domain, critical for stage differentiation has occurred in the D7B12 line (Shiels et al., 1998). Evidence has been found for epigenetic regulation of stage differentiation in a range of apicomplexan parasites, including *Plasmodium* and *Toxoplasma* (Bougdour et al., 2009), and the recent model proposed for commitment to gametocytogenesis incorporates chromatin modification via acetylation status (Chaal et al., 2010; Kafsack et al., 2014). With regard to this model and merogony in *Theileria*, it is highly significant that ApiAP2 domain proteins are associated with either recruitment or predicted histone acetylase/deacetylase activity (Lopez-Rubio et al., 2009; Salcedo-Amaya et al., 2009; Chaal et al., 2010). Investigation for motif/factor point mutations would require sequencing of the genome or

selected regions of the D7B12 genome. This approach was used by Sinha et al., (2014) for discovery of AP2-G but here the loss of differentiation potential was absolute not quantitative.

The stochastic model of *T. annulata* stage differentiation predicts the existence of auto-regulation of factors critical for commitment (Shiels et al., 1994). The presence of seven copies of CACACA/ACACAC motif in the upstream region of TA11145 strongly indicates that this ApiAP2 binds to its own promoter region and is auto-regulated. Despite this being a common motif of intergenic regions this level of occurrence is likely to be of functional significance, again supporting involvement of the TA11145 ApiAP2 in the differentiation process. Indeed, relative to the other (A)CACAC(A) motif binding ApiAP2s, the number of motifs upstream of TA11145 is greater. A full CACACA/ACACAC motif was found in only one copy in the upstream region of TA07100 and was not identified in IGR of TA19920. For TA07550 one ACACAC motif and a truncated GTCACA were detected on the sense strand and an ACACA and CACAC motif on the reverse.

The multiple presence of sites bound by an ApiAP2 in its own upstream region have also been found for AP2-G in *P. falciparum*, but the function of this arrangement remains to be determined. They may be required for higher affinity binding and for homo- or hetero-dimer formation and recruitment, as described for *Plasmodium* ApiAP2 domains (Lindner et al., 2010; Bougdour et al., 2010). One possibility is that it allows a quantitative increase in factor occupancy leading to an increase in factor levels. A second possibility is that it allows for occupancy of this promoter region by multiple factors binding to related motifs. The identification of five factors potentially binding (A)CACAC(A) motif in the genome of *T. annulata* with evidence of overlapping expression at the mRNA level suggests such a scenario could operate and supports a possible functional overlap of, or competition between, factors (Shiels et al., 1994). Gene expression in one stage could be then controlled by factors from a preceding or post stage that have differential affinity to the (A)CACAC(A)-motifs. Thus, although there is a high level of domain conservation between these ApiAP2 genes, differences between particular domains are visible, suggesting possible divergence in their optimally recognised target motifs. In this regard, it may be

of relevance that in the upstream region of TA11145 motif representation is: ACACAC x3; ACACACA x1 and CACACAC x1 and CACACA x1 (reverse strand), with one further ACACAC motif for *T. annulata* not conserved in *T. parva*.

Finally, the confirmation of CACACAC-binding by TA11145 AP2 domain fusion protein was achieved by EMSA. In addition, it was demonstrated that the domain binds to the double ACACAC/CACACA motif from its own upstream region suggesting that expression of the TA11145 gene involves auto-regulation by itself and/or other (A)CACAC(A)-binding factors. The possibility of multiple factor binding to this double ACACAC/CACACA motif was confirmed by EMSA with parasite enriched nuclear extracts from macroschizont infected (Day 0) and differentiating D7 cultures. Four potential band complexes (A-D, see Figure 4.32) were detected, although complex D was not observed in all EMSA experiments. Complex B was only detected in EMSA with PNE from differentiating parasite-infected cells while the two slower complexes C and D showed a distinct reduction using PNE derived from Day 7 and Day 9 D7 cultures. Specificity of binding of all four complexes was indicated by mutagenesis of the double motif. Moreover, the mutagenesis of motif 2 in isolation indicated that complex B showed differential specificity for the sequence ACACAC relative to complex C, or that the other complexes required the presence of both motifs. EMSA was also performed for the double ACACAC/CACACA motif with PNE from Day 0 (macroschizont) D7 compared to D7B12 cultures. As seen in Fig 4.36 only two of the specific complexes A and C were detected in D7, while for D7B12 only complex C was observed. This result indicates that there is a difference in expression of factors that bind this motif in cells that are competent or attenuated for *T. annulata* stage differentiation.

Clearly further work is required to identify the factors responsible for the native EMSA complexes, and validate that these complexes are conferred by the (A)CACAC(A) binding ApiAP2 factors identified in this study. Taking the accumulated results, it would seem most likely that the up-regulated complex B involves the TA11145 ApiAP2 domain based on its confirmed expression pattern. In a similar fashion shift A and possibly C can be associated with TA07550 based on the significantly reduced level of expression in the D7B12 line relative to D7. For the down-regulated complexes any of the other three (A)CACAC(A) AP2

domain proteins (and possibly TA07550) could be predicted and generation of specific antibodies could put more light on this subject. However, the data clearly indicates potential for a mechanism based on competition for motifs by factors with differential regulatory function, followed by a change-over to a regulatory factor governed by an auto-regulatory circuit.

Chapter 5

General discussion

5 General discussion

How a group of regulatory proteins accomplishes the change-over in expression of *Theileria* stage-specific gene sets was an interesting question and required both bioinformatics and molecular approaches to identify potential regulatory components of this process. Evidence had been generated previously allowing postulation that the mechanism that controls certain differentiation steps in Apicomplexa is conserved between members of this phylum, and may represent a primitive cellular differentiation mechanism retained by lower and higher eukaryotes (Shiels et al., 1997). Moreover, since the discovery of the ApiAP2 proteins by Balaji et al. (2008), several have been experimentally identified as stage-specific regulators of gene expression and commitment during the *Plasmodium* life cycle (Iyer et al., 2008; Yuda et al., 2009, Yuda et al., 2010; Iwanaga et al., 2012; Kafsack et al., 2014, Sinha et al., 2014), while others have been found to act as DNA-tethering proteins involved in control of heterochromatin dynamics (Flueck et al., 2010). Likewise, a study on *T. gondii* has revealed an ApiAP2 acting as a repressor of bradyzoite development (Radke et al., 2013). Therefore, it was not surprising that ApiAP2 proteins were identified as strong candidates to operate in the molecular mechanism responsible for regulating a key stage differentiation step in *Theileria*.

In this thesis, I present evidence that *T. annulata* ApiAP2s act as DNA binding proteins and are likely to function in transcriptional regulation. The data supported and expanded upon the hypothesis of a stochastic regulatory mechanism that controls the *T. annulata* macroschizont to merozoite transition (Shiels et al., 1994). Thus, analysis of *T. annulata* transcriptome data and bioinformatics revealed the genes encoding four ApiAP2 factors are up-regulated during merogony. The TaApiAP2 domains in these predicted polypeptides were conserved across three *Theileria* species and had orthologues either in *Plasmodium*, *Toxoplasma* or *Cryptosporidium* species. EMSA performed in Chapter 4 showed that they bind DNA motifs similar to those described for *P. falciparum* (Campbell et al., 2010), work in Chapter 3 showed that such motifs are over-represented in upstream regions of co-expressed gene clusters and that there is correlation between expression of AP2 domain genes and their putative targets.

While ApiAP2 orthologues may not operate in the same stage differentiation events, data generated in this study produced interesting correlation across *Theileria* and *Plasmodium*, suggesting that some commitment events could be conserved. In *Plasmodium* and *Theileria* a set of multiple AP2 domain factors bind (or can be predicted to bind to) (A)CACAC(A) motifs. These motifs are enriched in genes expressed during merogony. Moreover, the ApiAP2 factors predicted to bind the motif show a similar, overlapping expression pattern, with TA11145 clearly up-regulated. This data indicates that multiple factors could bind to the same motif and the same factor can bind to variant motif types. Both possibilities seem to be valid: four potential complexes were found to bind to the (A)CACAC(A) double motif upstream of TA11145, and it has been found that *Plasmodium* recombinant domains bind a primary motif but also recognise alternatives (Campbell et al., 2010). Based on the upstream region of the TA11145 gene a modified stochastic model for merogony, incorporating competitive binding to (A)CACAC(A) motifs can be postulated. This model assumes that auto-regulation of ApiAP2 gene-TA11145 is critical for commitment. Expression of the gene occurs at a low level in the macroschizont stage (based on array data) due to binding of (A)CACAC(A) motifs in the promoter by macroschizont ApiAP2 factors, with competition between factors that promote low (or repress) or high (or activate) level expression of the gene. When cells are placed at 41 °C an increase in protein levels relative to DNA generates a competitive advantage for up-regulation of TA11145 via preferential binding of its AP2 domain to multiple (A)CACAC(A) motifs in its own promoter (auto-regulation). A prediction of the model is that baseline expression and up-regulation of a merozoite regulator would be reduced in cells attenuated for the stage differentiation step, making the commitment step more difficult to attain. Two (A)CACAC(A) binding ApiAP2 factors (TA11145 and TA07550) fit this profile relative to the other three. Thus, initially low-level expression of TA11145 might involve TA07550 competing with other (A)CACAC(A) binding ApiAP2 factors, while high level expression only occurs following transition to full auto-regulation by TA11145 binding. One possibility is that this competitive binding operates to dictate factor accessibility by modulation of heterochromatin, with different ratios of competitors established between lines competent or attenuated for stage differentiation. Recent studies confirmed that *P.*

falciparum and *Toxoplasma* ApiAP2s are involved in heterochromatin maintenance (Flueck et al., 2010; Melo et al., 2013). It is likely that ApiAP2 factors are also involved in organization of heterochromatin in *Theileria*. Changes in expression levels of these ApiAP2(s) could affect not only their own expression, if they are auto-regulated, but also expression of other TFs and ultimately a large number of target genes that determine cellular status. H3K9me3 gene repression marks (trimethylation of histone 3 lysine 9) were found in the Pfl1085w (AP2-G) locus (an orthologue of TA13515 gene - *T. annulata* ApiAP2) (Lopez-Rubio et al., 2009; Salcedo-Amaya et al., 2009) and inhibition of *P. falciparum* histone deacetylases by apidicin interfered with methylation of this histone and was associated with elevated expression of Pfl1085w (Chaal et al., 2010).

Significant experimental investigation of the validity of this model is required, but was beyond the scope of this study due to time constraint. Fusion protein domains of all putative (A)CACAC(A) need to be generated with motif binding affinity established and evidence of competitive binding generated. Direct evidence that the EMSA complexes generated by nuclear extracts are due to binding of the candidate ApiAP2 factors is also needed. This would require raising antibodies against the ApiAP2 proteins and performing supershift EMSA. Data has been produced that apidicin can alter the differentiation potential and stage specific expression profile in *T. annulata* (Katzner and Shiels, unpublished) but a link to ApiAP2 factors and histone modifying enzymes requires to be established.

The updated model for merogony predicts that inhibition of (A)CACAC(A)-binding factor(s) would influence merozoite formation and provide a logical target for control strategies of parasite infection/transmission. One possibly would be to employ Py-Im polyamide mimetics developed to block binding of proteins to DNA (Dervan and Edelson, 2003). A Py-Im polyamide inhibitor was clearly shown in Chapter 4 to block binding of the AP2 domain of TA13515 by EMSA (see Figure 4.19). A preliminary experiment using a PA designed to bind to ACACAC showed a possible reduction in the level of the up-regulated EMSA shift obtained with nuclear extracts from cultures undergoing merogony. Further work using a PA

with higher predicted affinity is required. A major challenge for this type of strategy is delivery of the PA into the parasite nucleus.

Recent studies suggest that *Plasmodium* AP2-G transcription factor essential for commitment to gametogenesis is auto-regulated by a positive feedback loop mechanism (Sinha et al., 2014; Kafsack et al., 2014). A clear orthologue (TA13515) is present in *Theileria* and *Babesia* with evidence of up-regulation from merozoite to piroplasm stage and potential for auto-regulation. This suggests that this factor is expressed just prior to production of gametocytes in *Theileria*, as the piroplasm is the form ingested by the tick that gives rise to sexual stages in the lumen of the gut (Melhorn and Schein, 1984). It is therefore possible that the mechanism that governs transition to the sexual phase of the life cycle is conserved across apicomplexan genera. It will be of significance to further investigate this ApiAP2 at the protein level and if possible, validate, its potential essential role for *Theileria* sexual differentiation.

Analysis of *T. annulata* transcriptome and bioinformatics screening revealed two gene families (TashAT and SVSP) down-regulated from macroschizont to merozoite, which were found to contain specific C-box motifs in their upstream regions (4C-box for TashAT and 5C-box for SVSPs). Only one TashAT/TpHN-like gene, orthologue of the up-regulated Tash-a (TA03110) and TP01_0621, has been identified in the genome of the non-transforming *T. orientalis*. Phylogenetic data indicates that this gene is an ancestral member of the *T. annulata* TashAT and *T. parva* TpHN clusters, as no TashAT orthologues were found in *Plasmodium* and *B. bovis* genomes (Hayashida et al., 2012). The TA03110 gene also has a core C-box motif in its upstream region but lacked the flanking nucleotides found in the enlarged motif identified for other members of the cluster. A direct orthologue of *Plasmodium* AP2 domain genes that bind a C-box was not identified, but the TA12015 domain is an orthologue of a *C. parvum* AP2 domain that binds this motif. However, EMSA conducted with the binding domain of TA12015 and a G-box probe was unsuccessful. An EMSA experiment with differentiating PNE confirmed the C-box motif as a DNA binding site, however it also showed that nuclear factor binding to the C-box motif is down-regulated during merogony, suggesting it is unlikely to represent TA12015. In some members of the TashAT cluster the C-box motif is flanked by a TCTA motif

predicted to be recognised by the *Plasmodium* orthologue of TA16485. EMSA with a probe representing the upstream region of Tash1-like gene (TA03135) containing both the C-box and TCTA motifs gave clear shift with the TA16485 AP2 domain fusion protein suggesting that this TaApiAP2 may be involved in regulation of TashAT genes and, since it is up-regulated during merogony, would be likely to act as a repressor. A more complete understanding of the role the C-box motif plays in regulation of TashAT and SVSP gene expression is required, including identification of the down-regulated nuclear factor that binds to this motif. Regulation of these gene families could be associated with ability of *T. annulata* in establishment of the transformed infected leukocyte.

Presently, there is no proteomics data for the majority of the *T. annulata* life cycle. Examination of correlation between mRNA levels of TaApiAP2s and their respective proteins would provide information on the stages where they are most likely to function. Generation of specific ApiAP2 antibodies could confirm their cellular location during the differentiation event and assess whether changes observed at the mRNA levels are reflected in changes in the protein profile associated with merogony. Moreover, chromatin immunoprecipitation (ChIP) assay coupled with high-throughput next generation sequencing methodology could be used to investigate and map binding sites of *T. annulata* ApiAP2s and whether these are altered following a stage differentiation event or in cell lines with attenuated phenotypes.

Studies performed by Shiels et al. (1999) demonstrated that differentiation from the macroschizont to the merozoite stage is stochastic, however only limited information was available for *T. annulata* DNA motifs and factors that operate in this process. In *Plasmodium* ApiAP2 proteins have been predicted to form a regulator factor interaction network involving auto-regulation via their own promoter and recognition of other ApiAP2 gene promoters (Campbell et al., 2010). In *Theileria*, progression towards a quantitative commitment point has been proposed previously to operate via an auto-regulation event (Shiels et al., 1999) and this is supported by evidence that this occurs for TaApiAP2 genes up-regulated during merogony, particularly TA11145 with 7 potential sites of auto-regulation. In addition, like *Plasmodium*, regulation of TaApiAP2 expression may occur as a part of a cascade of factors during the differentiation event.

The identification of multiple motifs for ApiAP2 factor binding within the 4 up-regulated ApiAP2 identify potential forward and reverse (feedback) interactions together with auto-regulation in an ApiAP2 family regulation network (Figure 5.1) for *T. annulata*. As binding sites were only confirmed for TA13515, TA11145 and TA16485, TA12015 protein was excluded from the network. However, it is likely that other ApiAP2(s) and TFs are involved in such a regulator network; TA07750 and GATA, for example. Positive or negative regulatory feedback between ApiAP2s genes is possible. TA16485 is most likely to be the end point of the regulatory circuit, based on its expression profile (elevated late) and because it has least interactions predicted. The possibility that it acts downstream of TA11145 to down-regulate genes expressed in the macroschizont stage is a possibility, as it can bind to its motif in a TashAT gene IGR. It is probable that most of the ApiAP2s do not act in a master switch role, but are still important for defining a life cycle stage. In *P. berghei* a second ApiAP2 called AP2-G2 was identified as necessary although not essential for gametocytogenesis (Sinha et al., 2014).

The presence of predicted DNA motifs in upstream regions of regulator genes is not sufficient to explain the complex interplay with transcription factors that probably operates. The number and arrangement of motifs in upstream regions, orientation and spacing need to be taken into account. Multiple potential interactions between repressors and activators may occur. Moreover, some factors will be responsible for formation and maintenance of heterochromatin, as described for *P. falciparum* ApiAP2 (Flueck et al., 2010), and stabilisation of altered expression profiles.

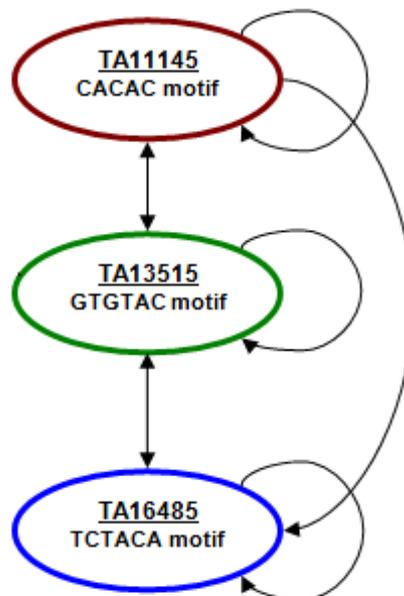


Figure 5.1. A regulatory network for the three macroschizont to merozoite up-regulated ApiAP2 genes. The network is based on the presence of AP2 domain target motifs in the upstream regions of the ApiAP2 genes. Arrows pointing away from ApiAP2 factors indicate regulation of other putative ApiAP2 target genes; auto-regulation is designated by the looped arrows. Positive or negative regulatory feedback between ApiAP2s is possible. TA16485 is most likely to be the end point of the regulatory circuit, as it has least interactions predicted.

In conclusion, work presented in the previous chapters of this thesis provides novel information on binding specificities of selected *T. annulata* ApiAP2 transcription factors and identifies them as candidate regulators of stage differentiation. Furthermore they point to potential commonality of differentiation steps across different Apicomplexa genera, support a stochastic differentiation model initially postulated by Shiels et al. (1999) and more recently suggested for *Plasmodium* gametocytogenesis (Kafsack et al., 2014). DNA binding factors conserved within the Apicomplexa performing functions critical for cellular survival or stage differentiation could provide targets for development of novel therapeutics. Thus blocking parasite progression from one stage to another could inhibit development of stages that generate pathology or allow transmission from mammalian host to vector. Based on this study and others, AP2-G and AP2 factors that bind the (A)CACAC(A) motifs are candidates for such a strategy. Further research is required to fully understand the functional role of ApiAP2 factors and how they interact with additional modulators of gene expression to regulate stage differentiation events.

References

Abrahamson, J.L., Lee, J.M., Bernstein, A. (1995) Regulation of p53-mediated apoptosis and cell cycle arrest by Steel factor. *Mol Cell Biol* 15(12):6953-60.

Adamson, R., Logan, M., Kinnaird, J., Langsley, G., and Hall, R. (2000) Loss of matrix metalloproteinase 9 activity in *Theileria annulata*-attenuated cells is at the transcriptional level and is associated with differentially expressed AP-1 species. *Mol Biochem Parasitol* 106: 51-61.

Allen, M.D., Yamasaki, K., Ohme-Takagi, M., Tateno, M., Suzuki, M. (1998) A novel mode of DNA recognition by a beta-sheet revealed by the solution structure of the GCC-box binding domain in complex with DNA. *EMBO J* 17:5484-96.

Allsopp, M.T. and Allsopp, B.A. (2006) Molecular sequence evidence for the reclassification of some *Babesia* species. *Ann N Y Acad Sci* 1081: 509-517.

Antia, R., Nowak, M.A., Anderson, R.M. (1996) Antigenic variation and the within-host dynamics of parasites. *Proc Natl Acad Sci U S A*. 93(3):985-9.

Ayoubi, T.A. and Van De Ven, W.J. (1996) Regulation of gene expression by alternative promoters. *FASEB J* 10(4):453-60. Review.

Bailey, T.L., Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2:28-36.

Bakheit, M.A. and Latif, A.A. (2002) The innate resistance of Kenana cattle to tropical theileriosis (*Theileria annulata* infection) in the Sudan. *Ann N Y Acad Sci* 969: 159-163.

Balaji, S., Babu, M.M., Iyer, L.M., Aravind, L. (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the

evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res* 33(13):3994-4006.

Balaji, S., Iyer, L.M., Babu, M.M., Aravind, L. (2008) Comparison of transcription regulatory interactions inferred from high-throughput methods: what do they reveal? *Trends Genet* 24(7):319-23.

Bansal, G.C. and Sharma, SP. (1986) Efficacy of parvaquone and long-acting oxytetracycline in *Theileria annulata* infection. *Vet Parasitol* 21(3):145-9.

Baumgartner, M., Chaussepied, M., Moreau, M.F., Werling, D., Davis, W.C., Garcia, A. (2000) Constitutive PI3-K activity is essential for proliferation, but not survival, of *Theileria parva*-transformed B cells. *Cell Microbiol* 2: 329-339.

Baylis, H.A., Megson, A., and Hall, R. (1995) Infection with *Theileria annulata* induces expression of matrix metalloproteinase 9 and transcription factor AP-1 in bovine leucocytes. *Mol Biochem Parasitol* 69: 211-222.

Behnke, M.S., Radke, J.B., Smith, A.T., Sullivan, W.J. Jr, White, M.W. (2008) The transcription of bradyzoite genes in *Toxoplasma gondii* is controlled by autonomous promoter elements. *Mol Microbiol* 68(6):1502-1518.

Behnke, M.S., Wootton, J.C., Lehmann, M.M., Radke, J.B., Lucas, O., Nawas, J., Sibley, L.D., White, M.W. (2010) Coordinated progression through two subtranscriptomes underlies the tachyzoite cycle of *Toxoplasma gondii*. *PLoS ONE* 5(8):e12354

Benjamini, Y., Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc Ser B (Methodological)* 57:289-300.

Benos, P.V., Bulyk, M.L., Stormo, G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucleic Acids Res* ;30(20):4442-51.

Berman, H.M., Battistuz, T., Bhat, T.N., Bluhm, W.F., Bourne, P.E., Burkhardt, K., Feng, Z., Gilliland, G.L., Iype, L., Jain, S., Fagan, P., Marvin, J., Padilla, D.,

Ravichandran, V., Schneider, B., Thanki, N., Weissig, H., Westbrook, J.D., Zardecki, C. (2002) The Protein Data Bank. *Acta Crystallogr D Biol Crystallogr* 58(Pt 6 No 1):899-907. Epub 2002 May 29.

Bettencourt, A., Franca, C., and Borges, J. (1907). Un cas de piroplasmose baciliforme chez le daim. *Institute Royal de Bacteriologie, Camara Pestana*, 1, 314-363.

Bishop, R., Gobright, E., Nene, V., Morzaria, S., Musoke, A., Sohanpal, B. (2000) Polymorphic open reading frames encoding secretory proteins are located less than 3 kilobases from *Theileria parva* telomeres. *Mol Biochem Parasitol* 110(2):359-71.

Blau, H.M. (1992) Differentiation requires continuous active control. *Annu Rev Biochem* 61:1213-30. Review.

Bohne, W., Parmley, S.F., Yang, S., Gross, U. (1996) Bradyzoite-specific genes. *Curr Top Microbiol Immunol* 219:81-91.

Bohne, W., Wirsing, A. and Gross, U. (1997). Bradyzoite-specific gene expression in *Toxoplasma gondii* requires minimal genomic elements. *Mol Biochem Parasitol* 85, 89-98.

Bouattour, A., Darghouth, M. A., and Ben, M. L. (1996) Cattle infestation by *Hyalomma* ticks and prevalence of *Theileria* in *H. detritum* species in Tunisia. *Vet Parasitol* 65: 233-245.

Bougdour, A., Braun, L., Cannella, D., Hakimi, M.A. (2010) Chromatin modifications: implications in the regulation of gene expression in *Toxoplasma gondii*. *Cell Microbiol* 1;12(4):413-23.

Boulter, N., and Hall, R. (1999) Immunity and vaccine development in the bovine theilerioses. *Adv Parasitol* 44: 41-97.

Branco, S., Orvalho, J., Leitão, A., Pereira, I., Malta, M., Mariano, I., Carvalho, T., Baptista, R., Shiels, B.R., Peleteiro, M.C. (2010) Fatal cases of *Theileria*

annulata infection in calves in Portugal associated with neoplastic-like lymphoid cell proliferation. *J Vet Sci* 11(1): 27-34.

Briegel, K., Bartunek, P., Stengl, G., Lim, K.C., Beug, H., Engel, J.D., Zenke, M. (1996) Regulation and function of transcription factor GATA-1 during red blood cell differentiation. *Development* 122(12):3839-50.

Brown, C.G. (1990) Control of tropical theileriosis (*Theileria annulata* infection) of cattle. *Parassitologia* 32: 23-31.

Brown, C.G.(1997). Dynamics and impact of tick-borne diseases of cattle. *Trop Anim Health Prod* 29(4 Suppl):1S-3S.

Bruce, M.C., Allano, P., Duthie, S., and Carter, R. (1990) Commitment of the malaria parasite *Plasmodium falciparum* to sexual and asexual development. *Parasitology* 100, 191-200.

Bustin SA. (2000) Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *J Mol Endocrinol.* 25: 169-193.

Butler, J., Kadonaga, J. (2002). The RNA Polymerase II core promoter: a key component to the regulation of gene expression. *Genes & Development* 16: 2583-2592.

Campbell, T.L., De Silva, E.K., Olszewski, K.L., Elemento, O., Llinás, M. (2010) Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathog* 6(10):e1001165.

Carroll, S.B. (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell* 134, 25-36

Carter, R., Miller, L.H. (1979) Evidence for environmental modulation of gametocytogenesis in *Plasmodium falciparum* in continuous culture. *Bull World Health Organ* 57 Suppl 1:37-52.

Chaal, B.K., Gupta, A.P., Wastuwidyaningtyas, B.D., Luah, Y.H., Bozdech, Z. (2010) Histone deacetylases play a major role in the transcriptional regulation of the *Plasmodium falciparum* life cycle. *PLoS Pathog* 22;6(1):e1000737.

Chaussepied, M., Janski, N., Baumgartner, M., Lizundia, R., Jensen, K., Weir, W., Shiels, BR., Weitzman, J.B., Glass, E.J., Werling, D., Langsley, G. (2010) TGF- β 2 induction regulates invasiveness of *Theileria*-transformed leukocytes and disease susceptibility. *PLoS Pathog* 6(11):e1001197.

Chen, X.M., Keithly, J.S., Paya, C.V., LaRusso, N.F. (2002). Cryptosporidiosis. *N Engl J Med* 346(22):1723-31.

Cock-Rada, A.M., Medjkane, S., Janski, N., Yousfi, N., Perichon, M., Chaussepied, M., Chluba, J., Langsley, G., Weitzman, J.B. (2012) SMYD3 promotes cancer invasion by epigenetic upregulation of the metalloproteinase MMP-9. *Cancer Res* 1;72(3):810-20. Epub 2011 Dec 22.

Cohn, B., Manque, P., Lara, A.M., Serrano, M., Sheth, N., Buck, G. (2010) Putative cis-regulatory elements associated with heat shock genes activated during excystation of *Cryptosporidium parvum*. *PLoS One* 5(3):e9512.

Coleman, B.I., Duraisingh, M.T. (2008) Transcriptional control and gene silencing in *Plasmodium falciparum*. *Cell Microbiol* (10):1935-46. Epub 2008 Jul 10.

Conrad, P. A., Kelly, B. G., and Brown, C. G. (1985) Intraerythrocytic schizogony of *Theileria annulata*. *Parasitology* 91 (Pt 1): 67-82.

Contreras, V.T., Morel, C.M., Goldenberg, S. (1985) Stage specific gene expression precedes morphological changes during *Trypanosoma cruzi* metacyclogenesis. *Mol Biochem Parasitol* 14(1):83-96.

Crabb, B.S., Cowman, A.F. (1996) Characterization of promoters and stable transfection by homologous and nonhomologous recombination in *Plasmodium falciparum*. *Proc Natl Acad Sci U S A* 93(14):7289-94.

Darghouth, M.A., Boulter, N.R., Gharbi, M., Sassi, L., Tait, A., Hall, R. (2006) Vaccination of calves with an attenuated cell line of *Theileria annulata* and the sporozoite antigen SPAG-1 produces a synergistic effect. *Vet Parasitol* 142(1-2):54-62.

Darkin-Rattray, S.J., Gurnett, A.M., Myers, R.W., Dulski, P.M., Crumley, T.M., Allocco, J.J., Cannova, C., Meinke, P.T., Colletti, S.L., Bednarek, M.A., Singh, S.B., Goetz, M.A., Dombrowski, A.W., Polishook, J.D., Schmatz, D.M. (1996) Apicidin: a novel antiprotozoal agent that inhibits parasite histone deacetylase. *Proc Natl Acad Sci U S A* 93(23):13143-7.

De Silva, E.K., Gehrke, A.R., Olszewski, K., León, I., Chahal, J.S., Bulyk, M.L., Llinás M. (2008) Specific DNA-binding by apicomplexan AP2 transcription factors. *Proc Natl Acad Sci U S A* 105(24):8393-8.

Deitsch, K., Duraisingh, M., Dzikowski, R., Gunasekera, A., Khan, S., Le Roch, K., Llinás, M., Mair, G., McGovern, V., Roos, D., Shock J., Sims, J., Wiegand, R., Winzeler, E. (2007) Mechanisms of gene regulation in *Plasmodium*. *Am J Trop Med Hyg* 77(2):201-8. Review.

Dermitzakis, E.T. and Clark, A.G. (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol* 19(7):1114-21.

Dervan, P.B., Edelson, B.S. (2003) Recognition of the DNA minor groove by pyrrole-imidazole polyamides. *Curr Opin Struct Biol*. 13(3):284-99.

D'Haeseleer, P. (2006) How does DNA sequence motif discovery work? *Nat Biotechnol* 24(8):959-61.

Dietz, K.J., Vogel, M.O., Viehhauser, A. (2010) AP2/EREBP transcription factors are part of gene regulatory networks and integrate metabolic, hormonal and environmental signals in stress acclimation and retrograde signalling. *Protoplasma* 245(1-4):3-14.

Dobbelaere, D. and Heussler, V. (1999) Transformation of leukocytes by *Theileria parva* and *T. annulata*. *Annu Rev Microbiol* 53: 1-42.

Dolan, T.T. (1989) Theileriasis: a comprehensive review. *Rev sci tech Off int Epiz* 8: 11-36.

Dolan, T.T., Young, A.S., Leitch, B.L., and Stagg, D.A. (1984) Chemotherapy of East Coast fever: parvaquone treatment of clinical disease induced by isolates of *Theileria parva*. *Vet Parasitol* 15: 103-116.

d'Oliveira, C., Feenstra, A., Vos, H., Osterhaus, A.D., Shiels, BR., Cornelissen A.W., Jongejan, F. (1997) Induction of protective immunity to *Theileria annulata* using two major merozoite surface antigens presented by different delivery systems. *Vaccine* 15(16); 1796-1804

Doniger, S.W., Huh, J., Fay, J.C. (2005) Identification of functional transcription factor binding sites using closely related *Saccharomyces* species. *Genome Res.* 15(5):701-9. Epub 2005 Apr 18.

Drews, G.N., Bowman, J.L., Meyerowitz, E.M. (1991) Negative regulation of the *Arabidopsis* homeotic gene *AGAMOUS* by the *APETALA2* product. *Cell* 65(6):991-1002.

Dschunkowsky, E. and Luhs, J. (1904) Die piroplasmosen der rinder, zentralblatt fur bacteriology. *Parasitenkunde, Infektionskrankheiten and Hygiene. Abteilung I. originale* 35,486-492.

Dumanli, N., Aktas, M., Cetinkaya, B., Cakmak, A., Koroglu, E., Saki, C.E. (2005) Prevalence and distribution of tropical theileriosis in eastern Turkey. *Vet Parasitol* 127: 9-15.

Durrani, Z., Weir, W., Pillai, S., Kinnaird, J., Shiels, B. (2012) Modulation of activation-associated host cell gene expression by the apicomplexan parasite *Theileria annulata*. *Cell Microbiol* (9):1434-54. Epub 2012 May 23.

Dyer, M, Tait, A. (1987) Control of lymphoproliferation by *Theileria annulata*. *Parasitol Today* 3(10):309-11.

Essien, K., Stoeckert, C.J. Jr (2010) Conservation and divergence of known apicomplexan transcriptional regulons. *BMC Genomics*. 2010 Mar 3;11:147.

Fawcett, D.W., Doxsey, S., Stagg, D.A. and Young, A.S. (1982). The entry of sporozoites of *Theileria parva* onto bovine lymphocytes in vitro - electron microscopic observations. *European Journal of Cell Biology* 27, 10-21.

Ferrer Rodríguez-Seguí, S., Akerman, I.J. (2012). GATA believe it: new essential regulators of pancreas development. *J Clin Invest* 122(10):3469-3471.

Flach, E.J. and Ouhelli, H. (1992) The epidemiology of tropical theileriosis (*Theileria annulata* infection in cattle) in an endemic area of Morocco. *Vet Parasitol* 44: 51-65.

Flueck, C., Bartfai, R., Niederwieser, I., Witmer, K., Alako, B.T., Moes, S., Bozdech, Z., Jenoe, P., Stunnenberg, H.G., Voss, T.S. (2010) A major role for the *Plasmodium falciparum* ApiAP2 protein PfSIP2 in chromosome end biology. *PLoS Pathog* 6(2):e1000784.

Forsyth, L.M., Minns, F.C., Kirvar, E., Adamson, R.E., Hall, F.R., McOrist, S. (1999) Tissue damage in cattle infected with *Theileria annulata* accompanied by metastasis of cytokine-producing, schizont-infected mononuclear phagocytes. *J Comp Pathol* 120: 39-57.

Fried MG (1989) Measurement of protein-DNA interaction parameters by electrophoresis mobility shift assay. *Electrophoresis* 10: 366 - 376

Garcia, C.R., de Azevedo, M.F., Wunderlich, G., Budu, A., Young, J.A., Bannister, L. (2008) *Plasmodium* in the postgenomic era: new insights into the molecular cell biology of malaria parasites. *Int Rev Cell Mol Biol* 266:85-156. Review.

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(5731):134-7.

Gardner, M.J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R.W., Carlton, J.M., Pain, A., Nelson, K.E., Bowman, S., Paulsen, I.T., James, K., Eisen, J.A., Rutherford, K., Salzberg, S.L., Craig A., Kyes, S., Chan, M.S., Nene, V., Shallom, S.J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M.W., Vaidya, A.B., Martin, D.M., Fairlamb, A.H., Fraunholz, M.J., Roos, D.S., Ralph, S.A., McFadden, G.I., Cummings, L.M., Subramanian, G.M., Mungall, C., Venter, J.C., Carucci, D.J., Hoffman, S.L., Newbold, C., Davis, R.W., Fraser, C.M., Barrell, B. (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419(6906):498-511.

Garner, M.M., Revzin, A. (1986) The use of gel electrophoresis to detect and study nucleic acid-protein interactions. *Trends Biol Sci* 11: 395 - 396

Gauer, M., Mackenstedt, U., Mehlhorn, H., Schein, E., Zapf, F., Njenga, E., Young, A., Morzaria, S. (1995) DNA measurements and ploidy determination of developmental stages in the life cycles of *Theileria annulata* and *T. parva*. *Parasitol Res* 81(7):565-74.

Gharbi, M., Sassi, L., Dorchies, P., and Darghouth, M.A. (2006) Infection of calves with *Theileria annulata* in Tunisia: Economic analysis and evaluation of the potential benefit of vaccination. *Vet Parasitol* 137: 231-241.

Gissot, M., Kim, K., Schaap, D., Ajioka, J.W. (2009) New eukaryotic systematics: a phylogenetic perspective of developmental gene expression in the Apicomplexa. *Int J Parasitol* 39(2):145-51.

Gissot, M., Choi, S., Thompson, R., Grealley, J.M., and Kim, K. 2008. *Toxoplasma gondii* and *Cryptosporidium parvum* lack detectable DNA cytosine methylation. *Eukaryot Cell*, 7(3): 537-540.

Glass, E.J. (2001) The balance between protective immunity and pathogenesis in tropical theileriosis: what we need to know to design effective vaccines for the future. *Res Vet Sci* 70: 71-75.

Glass, E.J., Innes, E.A., Spooner, R.L., and Brown, C.G. (1989) Infection of bovine monocyte/macrophage populations with *Theileria annulata* and *Theileria parva*. *Vet Immunol Immunopathol* 22: 355-368.

Glass, E.J., Preston, P.M., Springbett, A., Craigmile, S., Kirvar, E., Wilkie, G. *et al.* (2005) *Bos taurus* and *Bos indicus* (Sahiwal) calves respond differently to infection with *Theileria annulata* and produce markedly different levels of acute phase proteins. *Int J Parasitol* 35: 337-347.

Griffiths, R.B. and McCosker. (1990) Proceedings of the FAO expert consultation on revision of strategies for the control of ticks and tickborne diseases. *Parassitologia* 32, 1-209.

Gross, U., Bohne, W., Soete, M. and Dubremetz, J.F. (1996). Developmental differentiation between tachyzoites and bradyzoites of *Toxoplasma gondii*. *Parasitol Today* 12, 30-33.

Gubbels, M.J., d'Oliveira, C., and Jongejan, F. (2000) Development of an indirect Tams1 enzyme-linked immunosorbent assay for diagnosis of *Theileria annulata* infection in cattle. *Clin Diagn Lab Immunol* 7: 404-411.

Gubbels, M.J., Viseras, J., Habela, M.A., Jongejan, F.(2000B) Characterization of attenuated *Theileria annulata* vaccines from Spain and the Sudan. *Ann N Y Acad Sci* 916:521-32.

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.

Hakimi, M.A., Deitsch, K.W. (2007) Epigenetics in Apicomplexa: control of gene expression during cell cycle progression, differentiation and antigenic variation. *Curr Opin Microbiol* 10(4):357-62.

Hall, R., Ilhan, T., Kirvar, E., Wilkie, G., Preston, P.M., Darghouth, M., Somerville, R., Adamson, R. (1999) Mechanism(s) of attenuation of *Theileria annulata* vaccine cell lines. *Trop Med Int Health* 4(9):A78-84.

Haque, M., Singh, N. K., and Rath, S.S. (2010) Prevalence of *Theileria annulata* infection in *Hyalomma anatolicum anatolicum* in Punjab state, India. *J Parasit Dis* 34(1): 48-51.

Hawa, N., Rae, D.G., Younis, S., Mahadi, W., Ibrahim, R., and al-Wahab, W. (1988) Efficacy of parvaquone in the treatment of naturally occurring theileriosis in cattle in Iraq. *Trop Anim Health Prod* 20: 130-136.

Hayashida, K., Hara, Y., Abe, T., Yamasaki, C., Toyoda, A., Kosuge T., Suzuki, Y., Sato, Y., Kawashima, S., Katayama, T., Wakaguri, H., Inoue, N., Homma, K., Tada-Umezaki, M., Yagi, Y., Fujii, Y., Habara, T., Kanehisa, M., Watanabe, H., Ito, K., Gojobori, T., Sugawara, H., Imanishi, T., Weir, W., Gardner, M., Pain, A., Shiels, B., Hattori, M., Nene, V., Sugimoto, C. (2012) Comparative genome analysis of three eukaryotic parasites with differing abilities to transform leukocytes reveals key mediators of *Theileria*-induced leukocyte transformation. *MBio* 3(5):e00204-12.

Hertz-Fowler, C., Peacock, C.S., Wood, V., Aslett, M., Kerhornou, A., Mooney, P., Tivey, A., Berriman, M., Hall, N., Rutherford, K., Parkhill, J., Ivens, A.C., Rajandream, M.A., Barrell, B. (2004) GeneDB: a resource for prokaryotic and eukaryotic organisms. *Nucleic Acids Res* 32 (Database issue):D339-43.

Heussler, V.T., Rottenberg, S., Schwab, R., Kuenzi, P., Fernandez, P.C., McKellar, S. (2002) Hijacking of host cell IKK signalosomes by the transforming parasite *Theileria*. *Science* 298: 1033-1036.

Higuchi, R., Fockler, C., Dollinger, G., Watson, R. (1993) Kinetic PCR analysis: real-time monitoring of DNA amplification reactions. *Biotechnology* (N Y). 11(9):1026-30.

Hoke, S.M., Genereaux, J., Liang, G., Brandl, C.J.(2008) A conserved central region of yeast Ada2 regulates the histone acetyltransferase activity of Gcn5 and interacts with phospholipids. *J Mol Biol* 384(4):743-55. Epub 2008 Oct 11.

Hong, J.W., Hendrix, D.A., Levine, M.S. (2008) Shadow enhancers as a source of evolutionary novelty. *Science* 321(5894):1314.

Horrocks, P., Dechering, K., Lanzer, M. (1998) Control of gene expression in *Plasmodium falciparum*. *Mol Biochem Parasitol* 95:171-181.

Hu, J., Li, B., Kihara, D. (2005) Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res* 33(15):4899-913.

Hughes, J.D., Estep, P.W., Tavazoie, S., Church, G.M. (2000) Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J Mol Biol* 296(5):1205-14.

Hulliger, L. (1965) Cultivation of three species of *Theileria* in lymphoid cells in vitro. *J Protozool* 12(4):649-55.

Hulliger, L., Brown, C.G., Wilde, J.K. (1966) Transition of developmental stages of *Theileria parva* in vitro at high temperature. *Nature* 211(5046):328-9.

Hullinger, L., Wilde, K.H., Brown, C.G., Turner, L. (1964) Mode of multiplication of *Theileria* in cultures of bovine lymphocytic cells. *Nature* 203:728-30.

Iengar, P., Joshi, N.V. (2009) Identification of putative regulatory motifs in the upstream regions of co-expressed functional groups of genes in *Plasmodium falciparum*. *BMC Genomics* 10:18.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4(2):249-64.

Irvin, A.D. and Morrison, W. (1987) Immunopathology, immunology, and immunoprophylaxis of *Theileria* infections, in Immune responses in parasitic

infections: Immunology, Immunopathology, and immunoprophylaxis of *Theileria* infections. *E J L Soulsby, CRC Press Inc, Baton Rouge, Florida* 223-274.

Irvin, A.D., Cunningham, A.P. and Young, A.S. (1981) Advances in the Control of Theileriosis: International Conference Proceedings (Current Topics in Veterinary Medicine). *Mertinus Nijhoff Publishers, The Hague* 73-175.

Islam, M.K., Alim, M.A., Tsuji, N., and Mondal, M.M. (2006) An investigation into the distribution, host-preference and population density of ixodid ticks affecting domestic animals in Bangladesh. *Trop Anim Health Prod* 38: 485-490.

Islam, M.K., Jabbar, A., Campbell, B.E., Cantacessi, C., Gasser, R.B. (2011) Bovine theileriosis--an emerging problem in south-eastern Australia? *Infect Genet Evol* 11(8):2095-7. Epub 2011 Aug 31.

Iwanaga, S., Kaneko, I., Kato, T., Yuda, M. (2012) Identification of an AP2-family protein that is critical for malaria liver stage development. *PLoS One* 7(11):e47557. Epub 2012 Nov 7.

Iyer, L.M., Anantharaman, V., Wolf, M.Y., Aravind, L. (2008) Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. *Int J Parasitol* 38: 1-31.

Jacobson, S., Pillus, L. (2009) The SAGA subunit Ada2 functions in transcriptional silencing. *Mol Cell Biol* 29(22):6033-45. Epub 2009 Sep 8.

Jarrett, W.F., Crichton, G.W., Pirie, H.M. (1969) *Theileria parva*: kinetics of replication. *Exp Parasitol* 24(1):9-25.

Jensen, K., Talbot, R., Paxton, E., Waddington, D., Glass, E.J. (2006) Development and validation of a bovine macrophage specific cDNA microarray. *BMC Genomics* 7:224.

Jeyebal, L., Binod, K., Debdatta, R., Palavesam, A., Srikanta, G. (2012) Vaccine potential of recombinant antigens of *Theileria annulata* and *Hyalomma anatolicum anatolicum* against vector and parasite. *Veterinary Parasitology* 188(3-4):231-238

Jofuku, K.D., den Boer, B.G., Van Montagu, M., Okamoto, J.K. (1994) Control of *Arabidopsis* flower and seed development by the homeotic gene APETALA2. *Plant Cell* 6(9):1211-25.

Johnson, P.J., Kooter, J.M., Borst, P. (1987) Inactivation of transcription by UV irradiation of *T. brucei* provides evidence for a multicistronic transcription unit including a VSG gene. *Cell* 51(2):273-81.

Kafsack, B.F., Rovira-Graells, N., Clark, T.G., Bancells, C., Crowley, V.M., Campino, S.G., Williams, A.E., Drought, L.G., Kwiatkowski, D.P., Baker, D.A., Cortés, A., Llinás, M. (2014) A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature* 507(7491):248-52. Epub 2014 Feb 23.

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., Lander, E.S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423(6937):241-54.

Kent, W.J. (2002) BLAT-the BLAST-like alignment tool. *Genome Res* (4):656-64.

Kim, J., Klooster, S., Shapiro, D.J. (1995) Intrinsically bent DNA in a eukaryotic transcription factor recognition sequence potentiates transcription activation. *J Biol Chem* 270(3):1282-8.

Kinnaird, J.H., Weir, W., Durrani, Z., Pillai, S.S., Baird, M., Shiels, B.R. (2013) A Bovine Lymphosarcoma Cell Line Infected with *Theileria annulata* Exhibits an Irreversible Reconfiguration of Host Cell Gene Expression. *PLoS One* 8(6):e66833.

Kissinger, J.C., DeBarry, J. (2011) Genome cartography: charting the apicomplexan genome. *Trends Parasitol* 27(8):345-54

Kissinger, J.C., Gajria, B., Li L., Paulsen, I.T., Roos, D.S. (2003) ToxoDB: accessing the *Toxoplasma gondii* genome. *Nucleic Acids Res* 31(1):234-6.

Kundaje, A., Lianoglou, S., Li, X., Quigley, D., Arias, M., Wiggins, C.H., Zhang, L., Leslie, C. (2007) Learning regulatory programs that accurately predict differential expression with MEDUSA. *Ann N Y Acad Sci* 1115:178-202.

LaCount, D.J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J.R., Schoenfeld, L.W., Ota, I., Sahasrabudhe, S., Kurschner, C., Fields, S., Hughes, R.E. (2005) A protein interaction network of the malaria parasite *Plasmodium falciparum*. *Nature* 438(7064):103-7.

Latchman, D.S. (1991). Eukaryotic transcription factors. London: Academic Press.

Lau, A.O. (2009) An overview of the *Babesia*, *Plasmodium* and *Theileria* genomes: a comparative perspective. *Mol Biochem Parasitol* 164(1):1-8. Epub 2008 Dec 6. Review.

Lee, J.H., Cho, Y.S., Yoon, H.S., Suh, M.C., Moon, J., Lee, I., Weigel, D., Yun, C.H., Kim, J.K. (2005) Conservation and divergence of FCA function between *Arabidopsis* and rice. *Plant Molecular Biology* 58:823-838.

Leemans, I., Hooshmand-Rad, P., Brown, C.G., Kirvar, E., Wilkie, G., Uggla, A. (1998) In vitro infectivity and in vivo cross-protectivity of *Theileria lestoquardi* and *T. annulata* in sheep and cattle. *Ann N Y Acad Sci* 849:408-11.

Levine, N.D. (1985) Apicomplexa: The Piroplasms. *Veterinary Protozoology, The Iowa State University Press, Ames, Iowa*, p. 414.

Lim, L. and McFadden, G.I. (2009) The evolution, metabolism and functions of the apicoplast. *Philos Trans R Soc Lond B Biol Sci.* 365(1541): 749-763

Lindner, S.E., De Silva, E.K., Keck, J.L., Llinás, M. (2010) Structural determinants of DNA binding by a *P. falciparum* ApiAP2 transcriptional regulator. *J Mol Biol* 395(3):558-67

Lizundia, R., Chaussepied, M., Huerre, M., Werling, D., Di Santo, J.P., Langsley, G. (2006) c-Jun NH2-terminal kinase/c-Jun signaling promotes survival and metastasis of B lymphocytes transformed by *Theileria*. *Cancer research* 66(12):6105-10.

Lodish, H.F., Berk, A., Zipursky, L., Matsudaira, P., Baltimore, D. and Darnell, J.E. (1999) *Molecular Cell Biology*, 4th ed. Scientific American Press, N.Y.

Lopez-Rubio, J.J., Mancio-Silva, L., Scherf, A. (2009) Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. *Cell Host Microbe* 5(2):179-90.

Lu, W. and Yin, H. (1994). Bovine and ovine theileriosis in China and its immune prophylaxis. In: Spooner, R., Campbell, J., eds. Proceedings of the European Third Coordination Meeting on Tropical Theileriosis, 1994, Antalya, Turkey. The Roslin Institute (Edinburgh), Scotland, UK, 13-17.

MacHugh, N.D., Weir, W., Burrells, A., Lizundia, R., Graham, S.P., Taracha, E.L., Shiels, B.R., Langsley, G., Morrison, W.I. (2011) Extensive polymorphism and evidence of immune selection in a highly dominant antigen recognized by bovine CD8 T cells specific for *Theileria annulata*. *Infect Immun* 79(5):2059-69.

Magnani, E., Sjölander, K., Hake, S. (2004) From endonucleases to transcription factors: evolution of the AP2 DNA binding domain in plants. *Plant Cell* 16(9):2265-77. Epub 2004 Aug 19.

Mair, G.R., Braks, J.A., Garver, L.S., Wiegant, J.C., Hall, N., Dirks, R.W., Khan, S.M., Dimopoulos, G., Janse, C.J., Waters, A.P. (2006) Regulation of sexual development of Plasmodium by translational repression. *Science* 313(5787):667-9.

Mallick, K.P., Dhar, S., Malhotra, D.V., Bhushan, C., and Gautam, O.P. (1987) Immunization of neonatal bovines against *Theileria annulata* by an infection and treatment method. *Vet Parasitol* 24: 169-173.

Mantel, P.Y., Hoang, A.N., Goldowitz, I., Potashnikova, D., Hamza, B., Vorobjev, I., Ghiran, I., Toner, M., Irimia, D., Ivanov, A.R. (2013) Malaria microvesicles act as messengers that induce potent responses in immune cells and parasites. *Cell Host Microbe* 13: 521-534

McDonald, V., Rose, M.E. (1987) *Eimeria tenella* and *E. necatrix*: a third generation of schizogony is an obligatory part of the developmental cycle. *J Parasitol* 73(3):617-22.

McGuffin, L.J., Bryson, K., Jones, D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics* 16(4):404-5.

McHardy, N., Morgan, D.W. (1985) Treatment of *Theileria annulata* infection in calves with parvaquone. *Res Vet Sci* 39(1):1-4.

McHardy, N., Wekesa, L.S., Hudson, A.T., and Randall, A.W. (1985) Antitheilerial activity of BW720C (buparvaquone): a comparison with parvaquone. *Res Vet Sci* 39: 29-33.

McRobert, L., Taylor, C.J., Deng, W., Fivelman, Q.L., Cummings, R.M. (2008) Gametogenesis in malaria parasites is mediated by the cGMP-dependent protein kinase. *PLoS Biol* 6(6): e139.

Mehlhorn, H. and Raether, W. (1988) Effects of halofuginone lactate on lymphocytes and MDBK cells infected with *Theileria annulata* and/or vesicular stomatitis virus (VSV): an in vitro study. *Parasitol Res* 74(5):441-7.

Mehlhorn, H. and Shein, E. (1984) The piroplasms: life cycle and sexual stages. *Adv Parasitol* 23: 37-103.

Meissner, M., Soldati, D. (2005) The transcription machinery and the molecular toolbox to control gene expression in *Toxoplasma gondii* and other protozoan parasites. *Microbes Infect* 7(13):1376-84. Epub 2005 Jun 29. Review.

Melo, M.B., Nguyen, Q.P., Cordeiro, C., Hassan, M.A., Yang, N., McKell, R., Rosowski, E.E., Julien, L., Butty, V., Dardé, M.L., Ajzenberg, D., Fitzgerald, K., Young, L.H., Saeij, J.P. (2013) Transcriptional analysis of murine macrophages infected with different *Toxoplasma* strains identifies novel regulation of host signaling pathways. *PLoS Pathog* 9(12):e1003779. Epub 2013 Dec 19.

Meng, K., Li, Z., Wang, Y., Jing, Z., Zhao, X., Liu, J., Cai, D., Zhang, L., Yang, D., Wang, S. (2014) PCR-based detection of *Theileria annulata* in *Hyalomma*

asiaticum ticks in northwestern China. *Ticks Tick Borne Dis* 5(2):105-6. Epub 2013 Oct 24.

Mhadhbi, M., Naouach, A., Boumiza, A., Chaabani, M.F., Abderazzak, S., Darghouth, M.A. (2010) In vivo evidence for the resistance of *Theileria annulata* to buparvaquone. *Vet Parasitol* 169(3-4):241-7. Epub 2010 Jan 25.

Mikolajczak, S.A., Silva-Rivera, H., Peng, X., Tarun, A.S., Camargo, N., Jacobs-Lorena, V., Daly, T.M., Bergman, L.W., de la Vega, P., Williams, J., Aly, A.S., Kappe, S.H. (2008) Distinct malaria parasite sporozoites reveal transcriptional changes that cause differential tissue infection competence in the mosquito vector and mammalian host. *Mol Cell Biol* (20):6196-207.

Militello, K.T., Dodge, M., Bethke, L., Wirth, D.F. (2004) Identification of regulatory elements in the *Plasmodium falciparum* genome. *Mol Biochem Parasitol* 134(1):75-88.

Morris, J.F., Hromas, R., Rauscher, F.J. 3rd. (1994) Characterization of the DNA-binding properties of the myeloid zinc finger protein MZF1: two independent DNA-binding domains recognize two DNA consensus sequences with a common G-rich core. *Mol Cell Biol* 14(3):1786-95.

Morrison W.I., McKeever D.J. (2006) Current status of vaccine development against *Theileria* parasites. *Parasitology* 133 Suppl:S169-87.

Morzaria, S., Nene, V., Bishop, R., Musoke, A. (2000) Vaccines against *Theileria parva*. *Ann N Y Acad Sci* 916:464-73.

Mullapudi, N., Joseph, S.J., Kissinger, J.C. (2009) Identification and functional characterization of cis-regulatory elements in the apicomplexan parasite *Toxoplasma gondii*. *Genome Biol* 10(4):R34. Epub 2009 Apr 7.

Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247(4):536-40.

- Myler, P.J., Beverley, S.M., Cruz, A.K., Dobson, D.E., Ivens, A.C. (2001) The *Leishmania* genome project: New insights into gene organization and function. *Med Microbiol Immunol (Berl)* 190: 9-12.
- Nassar, A.M. (1992) *Theileria* infection in camels (*Camelus dromedarius*) in Egypt. *Vet Parasitol* 43(1-2):147-9.
- Neitz, W.O. (1957). Theileriosis, gonderosis and cyauxzoonoses: a review. *Gonderia annulata* infection. *Onderstepoort Journal of Veterinary Research* 27, 319-326.
- Nene, V., Bishop, R., Morzaria, S., Gardner, M.J., Sugimoto, C., ole-MoiYoi, O.K., Fraser, C.M., Irvin, A. (2000) *Theileria parva* genomics reveals an atypical apicomplexan genome. *Int J Parasitol* 30(4):465-74.
- Ngumi, P.N., Young, A.S., Lampard, D., Mining, S.K., Ndungu, S.G., Lesan, A.C. (1992) Further evaluation of the use of buparvaquone in the infection and treatment method of immunizing cattle against *Theileria parva* derived from African buffalo (*Syncerus caffer*). *Vet Parasitol* 43: 15-24.
- Norval, R.A.I., Perry, B.D., and Young, A.S. (1992) The Epidemiology of Theileriosis in Africa. *London: Academic Press*.
- Oberstaller, J., Joseph, S.J., Kissinger, J.C. (2013) Genome-wide upstream motif analysis of *Cryptosporidium parvum* genes clustered by expression profile. *BMC Genomics* 29;14:516
- Ohyama, T. (2001) Intrinsic DNA bends: an organizer of local chromatin structure for transcription. *Bioessays* 23(8):708-15.
- Olobo, J.O., Black, S.J. (1989) Selected phenotypic and cloning properties of a bovine lymphoblastoid cell line, BL20. *Vet Immunol Immunopathol* 20(2):165-72.
- Ono, T., Ohnishi, Y., Nagamune, K., Kano, M. (1993) Gametocytogenesis induction by Berenil in cultured *Plasmodium falciparum*. *Exp Parasitol* 77(1):74-8.

Onuma, M., Kubota, S., Kakuda, T., Sako, Y., Sugimoto, C. (1997) Vaccine development against *Theileria* parasite. *Southeast Asian J Trop Med Public Health* 28 Suppl 1:148-54.

Osman, S.A., Al-Gaabary, M.H. (2007) Clinical, haematological and therapeutic studies on tropical theileriosis in water buffaloes (*Bubalus bubalis*) in Egypt. *Vet Parasitol* 146(3-4):337-40. Epub 2007 Apr 8.

Pain, A., Renauld, H., Berriman, M., Murphy, L., Yeats, C.A., Weir, W. (2005) Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. *Science* 309: 131-133.

Painter, H.J., Campbell, T.L., Llinás, M. (2011) The Apicomplexan AP2 family: integral factors regulating *Plasmodium* development. *Mol Biochem Parasitol* 176(1):1-7. Epub 2010 Nov 30.

Pérez-Martín, J, de Lorenzo, V. (1997) Clues and consequences of DNA bending in transcription. *Annu Rev Microbiol* 51:593-628. Review.

Pérez-Martín, J., Rojo, F., de Lorenzo, V. (1994) Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. *Microbiol Rev* 58(2):268-90.

Pipano, E., Samish, M., Kriegel, Y., Yeruham, I. (1981) Immunization of Friesian cattle against *Theileria annulata* by the infection-treatment method. *Br Vet J* 137(4):416-20.

Pipano, E., Shkap, V. (2000) Vaccination against tropical theileriosis. *Ann N Y Acad Sci* 916:484-500.

Pollastri, G., Przybylski, D., Rost, B., Baldi, P. (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* 47(2):228-35.

Preston, P.M., Brown, C.G., Bell-Sakyi, L., Richardson, W., and Sanderson, A. (1992) Tropical theileriosis in *Bos taurus* and *Bos taurus* cross *Bos indicus* calves:

response to infection with graded doses of sporozoites of *Theileria annulata*. *Res Vet Sci* 53: 230-243.

Puiu, D., Enomoto, S., Buck, G.A., Abrahamsen, M.S., Kissinger, J.C. (2004) CryptoDB: the Cryptosporidium genome resource. *Nucleic Acids Res* 1;32.

Purnell, R.E. (1978) *Theileria annulata* as a hazard to cattle in countries on the Northern Mediterranean Littoral. *Veterinary Science Communications* 2,3-11.

Radke, J.B., Lucas, O., De Silva, E.K., Ma, Y., Sullivan, W.J. Jr, Weiss, L.M., Llinas, M., White, M.W. (2013) ApiAP2 transcription factor restricts development of the *Toxoplasma* tissue cyst. *Proc Natl Acad Sci USA* 23;110(17):6871-6. Epub 2013 Apr 9.

Reeves, R. (2001) Molecular biology of HMGA proteins: hubs of nuclear function. *Gene* 277: 63-81.

Regev-Rudzki, N., Wilson, D.W., Carvalho, T.G., Sisquella, X., Coleman, B.M., Rug, M., Bursac, D., Angrisano, F., Gee, M., Hill, A.F., Baum, J., Cowman, A.F. (2013) Cell-cell communication between malaria-infected red blood cells via exosome-like vesicles. *Cell* 153(5):1120-33. Epub 2013 May 15.

Reid, G.D, Bell, L.J. (1984) The development of *Theileria annulata* in the salivary glands of the vector tick *Hyalomma anatolicum anatolicum*. *Ann Trop Med Parasitol* 78(4):409-21.

Reuner, B., Vassella, E., Yutzy, B., Boshart, M. (1997) Cell density triggers slender to stumpy differentiation of *Trypanosoma brucei* bloodstream forms in culture. *Mol Biochem Parasitol* 90(1):269-80.

Robinson, P. M. (1982) *Theileriosis annulata* and its transmission-a review. *Trop Anim Health Prod* 14: 3-12.

Rodríguez-Seguí, S., Akerman, I., Ferrer, J. (2012) GATA believe it: new essential regulators of pancreas development. *J Clin Invest* 122(10):3469-71.

Roos, D.S., Donald, R.G., Morrissette, N.S., Moulton, A.L. (1994) Molecular tools for genetic dissection of the protozoan parasite *Toxoplasma gondii*. *Methods Cell Biol* 45:27-63.

Rosenthal, P.J., McKerrow, J.H., Aikawa, M., Nagasawa, H., Leech, J.H. (1988) A malarial cysteine proteinase is necessary for hemoglobin degradation by *Plasmodium falciparum*. *J Clin Invest* 82,1560-1566.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A., Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* 16(10):944-5.

Sacks, D.L., Perkins, P.V. (1984) Identification of an infective stage of *Leishmania promastigotes*. *Science* 223(4643):1417-9.

Salcedo-Amaya, A.M, van Driel, M.A., Alako, B.T., Trelle, M.B., van den Elzen, A.M., Cohen, A.M., Janssen-Megens, E.M., van de Vegte-Bolmer, M., Selzer, R.R., Iniguez, A.L., Green, R.D., Sauerwein, R.W., Jensen, O.N., Stunnenberg, H.G. (2009) Dynamic histone H3 epigenome marking during the intraerythrocytic cycle of *Plasmodium falciparum*. *Proc Natl Acad Sci USA*. 106(24):9655-60.

Samantaray, S.N., Bhattacharyulu, Y. and Gill, B.S. (1980) Immunisation of calves against bovine tropical theileriosis (*Theileria annulata*) with graded doses of sporozoites and irradiated sporozoites. *Int J Parasitol* 10: 355-358.

Schaeffer, H.J., Weber, M.J. (1999) Mitogen-activated protein kinases: specific messages from ubiquitous messengers. *Mol Cell Biol* 19(4):2435-44.

Schein, E. and Friedhoff, K.T. (1978) Light microscopic studies on the development of *Theileria annulata* (Dschunkowsky and Luhs, 1904) in *Hyalomma anatolicum excavatum* (Koch, 1844). II. The development in haemolymph and salivary glands (author's transl)]. *Z Parasitenkd* 56: 287-303.

Schein, E. and Voigt, W.P. (1979) Chemotherapy of bovine theileriosis with Halofuginone. Short communication. *Acta Trop* 36: 391-394.

Schein, E., Buscher, G., and Friedhoff, K.T. (1975) Light microscopic studies on the development of *Theileria annulata* (Dschunkowsky and Luhs, 1904) in *Hyalomma anatolicum excavatum* (Koch, 1844). I. The development in the gut of engorged nymphs (author's transl). *Z Parasitenkd* 48: 123-136.

Schmuckli-Maurer, J., Casanova, C., Schmied, S., Affentranger, S., Parvanova, I. (2009) Expression Analysis of the *Theileria parva* Subtelomere-Encoded Variable Secreted Protein Gene Family. *PLoS ONE* 4(3): e4839.

Serfling, E. (1989) Autoregulation--a common property of eukaryotic transcription factors? *Trends Genet* 5(5):131-3.

Shaw, M. K., Tilney, L.G., and Musoke, A.J. (1991) The entry of *Theileria parva* sporozoites into bovine lymphocytes: evidence for MHC class I involvement. *J Cell Biol* 113: 87-101.

Shaw, M.K. (1997) The same but different: the biology of *Theileria* sporozoite entry into bovine cells. *Int J Parasitol* 27: 457-474.

Shaw, M.K., Tilney, L.G. (1992) How individual cells develop from a syncytium: merogony in *Theileria parva* (Apicomplexa). *Cell Sci* 101 (Pt 1):109-23.

Shiels, B. (1999) Should I stay or should I go now? A stochastic model of stage differentiation in *Theileria annulata*. *Parasitol Today* 15(6):241-5.

Shiels, B., Aslam, N., McKellar, S., Smyth, A., Kinnaird, J. (1997) Modulation of protein synthesis relative to DNA synthesis alters the timing of differentiation in the protozoan parasite *Theileria annulata*. *J Cell Sci* 110 (Pt 13):1441-51.

Shiels, B., Fox, M., McKellar, S., Kinnaird, J., Swan, D. (2000) An upstream element of the TamS1 gene is a site of DNA-protein interactions during differentiation to the merozoite in *Theileria annulata*. *J Cell Sci* 113 (Pt 12):2243-52.

Shiels, B., Kinnaird, J., McKellar, S., Dickson, J., Miled, L.B., Melrose, R. (1992) Disruption of synchrony between parasite growth and host cell division is a determinant of differentiation to the merozoite in *Theileria annulata*. *J Cell Sci* 101:99-107.

Shiels, B., Langsley, G., Weir, W., Pain, A., McKellar, S., and 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., Swan, D., McKellar, S., Aslam, N., Dando, C., Fox, M., Ben-Miled, L., Kinnaird, J. (1998) Directing differentiation in *Theileria annulata*: old methods and new possibilities for control of apicomplexan parasites. *Int J Parasitol* 28(11):1659-70.

Shiels, B.R., Kinnaird, J., Dickson, J., Tetley, L., McKellar, S., Glascodine, J., Brown, C.G.D., Tait, A. (1992) Investigations into regulation of stage differentiation in *Theileria annulata*. *Sci* 101 (Pt 1): 99-107.

Shiels, B.R., Smyth, A., Dickson, J., McKellar, S., Tetley, L., Fujisaki, K., Hutchinson, B., and Kinnaird, J.H. 1994. A stoichiometric model of stage differentiation in the protozoan parasite *Theileria annulata*. *Mol Cell Differ* 2:101-125.

Singh, A., Singh, J., Grewal, A.S., Brar, R.S.(2001) Studies on some blood parameters of crossbred calves with experimental *Theileria annulata* infections. *Vet Res Commun* 25(4):289-300.

Singh, D.K., Jagdish, S., Gautam, O.P., and Dhar, S. (1979) Infectivity of ground-up tick supernates prepared from *Theileria annulata* infected *Hyalomma anatolicum anatolicum*. *Trop Anim Health Prod* 11: 87-90.

Singh, J., Gill, J.S., Kwatra, M.S., and Sharma, K.K. (1993) Treatment of theileriosis in crossbred cattle in the Punjab. *Trop Anim Health Prod* 25: 75-78.

Sinha, A., Hughes, K.R., Modrzynska, K.K., Otto, T.D., Pfander, C., Dickens, N.J., Religa, A.A., Bushell, E., Graham, A.L., Cameron, R., Kafsack, B.F., Williams, A.E., Llinás, M., Berriman, M., Billker, O., Waters, A.P. (2014) A cascade of DNA-binding proteins for sexual commitment and development in *Plasmodium*. *Nature* 13;507(7491):253-7.

Soète, M., Camus, D., Dubremetz, J.F. (1994) Experimental induction of bradyzoite-specific antigen expression and cyst formation by the RH strain of *Toxoplasma gondii* in vitro. *Exp Parasitol* 78(4):361-70.

Spooner, R.L., Innes, E.A., Glass, E.J., and Brown, C.G. (1989) *Theileria annulata* and *T. parva* infect and transform different bovine mononuclear cells. *Immunology* 66: 284-288.

Stern, D.L., Orgogozo, V. (2008) The loci of evolution: how predictable is genetic evolution? *Evolution* 62(9):2155-77. Epub 2008 Jul 4.

Stockinger, E.J., Mao, Y., Regier, M.K., Triezenberg, S.J., Thomashow, M.F. (2001) Transcriptional adaptor and histone acetyltransferase proteins in *Arabidopsis* and their interactions with CBF1, a transcriptional activator involved in cold-regulated gene expression. *Nucleic Acids Res* 29(7):1524-33.

Stormo, G.D. and Fields, D.S. (1998) Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci* 23, 109-113.

Sudarsanam, P., Winston, F. (2000) The Swi/Snf family nucleosome-remodeling complexes and transcriptional control. *Trends Genet* 16(8):345-51.

Sugimoto, C., Kawazu, S., Kamio, T., Fujisaki, K. (1991) Protein analysis of *Theileria sergenti/buffeli/orientalis* piroplasms by two-dimensional polyacrylamide gel electrophoresis. *Parasitology* 102 Pt 3:341-6.

Sunil, S., Chauhan, V.S., Malhotra, P. (2008) Distinct and stage specific nuclear factors regulate the expression of falcipains, *Plasmodium falciparum* cysteine proteases. *BMC Mol Biol* 14;9:47.

Swan, D.G., Stadler, L., Okan, E., Hoffs, M., Katzer, F., Kinnaird, J. (2003) TashHN, a *Theileria annulata* encoded protein transported to the host nucleus displays an association with attenuation of parasite differentiation. *Cell Microbiol* 5: 947-956.

Tang, L., Nogales, E., and Ciferri, C. (2010) Structure and function of SWI/SNF chromatin remodeling complexes and mechanistic implications for transcription. *Prog Biophys Mol Biol* 102(2-3): 122-128. Published online May 20, 2010.

Tarun, A.S., Peng, X., Dumpit, R.F., Ogata, Y., Silva-Rivera, H., Camargo, N., Daly, T.M., Bergman, L.W., Kappe, S.H. (2008) A combined transcriptome and proteome survey of malaria parasite liver stages. *Proc Natl Acad Sci U S A* 105(1):305-10.

Taylor, M.A., Coop, R.L., and Wall, R.L. (2007) Parasites of cattle. In *Veterinary Parasitology, 3rd ed Blackwell, Oxford*, 93-151.

Thomas, M.C., Chiang, C.M. (2006) The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol.* 41(3):105-78. Review.

Tompa, M., Li N., Bailey, T.L., Church, G.M., De Moor, B., Eskin E., Favorov, A.V., Frith, M.C., Fu, Y., Kent, W.J., Makeev, V.J., Mironov, A.A., Noble, W.S., Pavese, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C., Zhu, Z. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol* 23(1):137-44.

Uilenberg, G. (1981) *Theileria* species of domestic livestock. In: Irvin, A.D., Cunningham, M.P., Young, A.S.(Eds.), *Advances in Control of Theileriosis*. Nijhoff (Martinus). *The Hague* 4-37.

Van der Ploeg, L.H., Giannini, S.H., Cantor, C.R. (1985) Heat shock genes: regulatory role for differentiation in parasitic protozoa. *Science* 228(4706):1443-6.

- von Schubert, C., Xue, G., Schmuckli-Maurer, J., Woods, K.L., Nigg, E.A., Dobbelaere, D.A. (2010) The transforming parasite *Theileria* co-opts host cell mitotic and central spindles to persist in continuously dividing cells. *PLoS Biol* 8(9). pii: e1000499.
- Walker, R., Gissot, M., Croken, M.M., Huot, L., Hot, D., Kim, K., Tomavo, S. (2013) The *Toxoplasma* nuclear factor TgAP2XI-4 controls bradyzoite gene expression and cyst formation. *Mol Microbiol* 87(3):641-655.
- Waller, R.F., McFadden, G.I. (2005) The apicoplast: a review of the derived plastid of apicomplexan parasites. *Curr Issues Mol Biol* 7(1):57-79.
- Wang, J., Dixon, S.E., Ting, L-M., Liu, T-K., Jeffers, V. (2014) Lysine Acetyltransferase GCN5b Interacts with AP2 Factors and Is Required for *Toxoplasma gondii* Proliferation. *PLoS Pathog* 10(1): e1003830.
- Weir, W., Karagenc, T., Baird, M., Tait, A., and Shiels, B.R. (2010) Evolution and diversity of secretome genes in the apicomplexan parasite *Theileria annulata*. *BMC Genomics* 11: 42.
- Wittkopp, P., Kalay, G. (2012) Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nature Review Genetics* 13: 59-69.
- Wray, G.A. (2007). The evolutionary significance of cis-regulatory mutations. *Nature Reviews Genetics* 8 (3): 206-216.
- Yin, H., Luo, J., Zhang, Q., Guan, G., Lu, B., Guo, J., Wang, Y. and Lu, W. (2003) Attempted Transmission of *Theileria annulata* to Sheep and Goats with *Hyalomma detritum* and *Hyalomma anatolicum*. *Journal of Animal and Veterinary Advances* 2(6):372-374.
- Young, A.S., Burridge, M.J., and Payne, R.C. (1977) Transmission of a *Theileria* species to cattle by the ixodid tick, *Amblyomma cohaerens* *Trop Anim Health Prod* 9: 37-45.

Young, A.S., Shaw, M.K., Ochanda, H., Morzaria, S.P., Dolan, T.T. (1992) Factors affecting the transmission of African *Theileria* species of cattle by ixodid ticks (conference proceedings). *Saint Paul, Minnesota* 15-18.

Young, J.A., Johnson, J.R., Benner, C., Yan, S.F., Chen, K., Le Roch, K.G., Zhou, Y., Winzeler, E.A. (2008) In silico discovery of transcription regulatory elements in *Plasmodium falciparum*. *BMC Genomics* 9:70.

Yuda, M., Iwanaga, S., Shigenobu, S., Kato, T., Kaneko, I. (2009) Identification of a transcription factor in the mosquito-invasive stage of malaria parasites. *Mol Microbiol* 71(6):1402-1414

Yuda, M., Iwanaga, S., Shigenobu, S., Kato, T., Kaneko, I. (2010) Transcription factor AP2-Sp and its target genes in malarial sporozoites. *Mol Microbiol* 75(4):854-863.

Zhang, Z. and Gerstein, M. (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucleic Acids Res* 31(18):5338-48

APPENDICES

Appendix 1 – Materials and methods

1.1. List of buffer and solution recipes commonly used in the experimental protocols of this thesis.

Phosphate buffered saline (PBS X 1)

Dissolve 8g NaCl, 0.2g KCl, 1.15g Na₂HPO₄ and 0.2g KH₂PO₄ in dH₂O (up to 1L of total volume) and adjust to pH 7.4.

DNA electrophoresis Gel

Dissolve 1g UltraPure Agarose in 100ml 1 X TAE, heat in microwave. Cool down and add 2-3µl of Ethidium Bromide.

Paraformaldehyde fixation solution (3.7 %)

Dissolve 3.7g paraformaldehyde in 100ml of 1 X PBS. Add 150µl of 1M NaOH, heat to 50°C in microwave (do not boil), cool on ice and adjust pH to 7.4 with HCl. Store at 4°C.

Permeabilisation solution (Triton X 0.2)

Mix 200µl Triton X-100 with 100ml 1 x PBS, stir continuously until dissolved and store at room temperature.

Mounting medium

For 1 ml solution, add 50% glycerol in 1 x PBS (i.e. 0.5ml glycerol, 0.5ml 1 X PBS), 10µl of anti-fade reagent (diphenylamine at 100mg/ml) and 1µl of DAPI (4', 6-diamidino-2-phenylindole, dihydrochloride) for nuclear staining. Store at -20°C before use (lasts for about 1 month).

0.1% Evans blue stain

Dissolve 0.1g Evans blue in 90ml 1 x PBS and bring to 100 ml with 1 X PBS. Store at room temperature.

10% SDS

Dissolve 10g SDS in 90 ml dH₂O and bring to 100 ml with dH₂O. Store at room temperature.

10x SDS-PAGE running buffer, pH 8.3 (1 L)

Dissolve 30.3 g Tris-base, 144.0 g Glycine, and 10.0 g SDS with dH₂O up to 1 L of total volume and store at 4 °C before use. Dilute 50 ml of 10x stock with 450 ml dH₂O for each electrophoresis.

Separating gel buffer (1.5 M Tris-HCl, pH 8.8)

Dissolve 18.2 g Tris-base in 80 ml dH₂O. Adjust to pH 8.8 with HCl. Bring total volume to 100 ml with dH₂O and store at 4 °C.

Stacking gel buffer (0.25 M Tris-HCl, 0.2 % SDS, pH 6.8)

Dissolve 3 g Tris base, 2 ml of 10 % SDS in 60 ml dH₂O. Adjust to pH 6.8 with HCl. Bring total volume to 100 ml with dH₂O and store at 4 °C.

10% ammonium persulfate (APS)

Dissolve 0.1 g ammonium persulfate (APS) in 1 ml dH₂O. Prepare fresh when required.

Separating gel formulation (10% SDS-PAGE gel)

For 10% resolving gel, mix 4.05 ml dH₂O, 3.3 ml 30% Acrylamide/Bis, 2.5 ml 1.5M Tris-HCl pH 8.8, 100 µl 10% SDS, 50 µl 10% APS and 5 µl TEMED.

5% Stacking gel formulation

Mix 1.8 ml dH₂O, 0.63 ml 30% Acrylamide/Bis, 2.5 ml of stacking gel buffer, 50 µl 10% SDS, 50 µl 10% APS and 5.0 µl TEMED.

Coomassie blue staining R-250 (500 ml)

Dissolve 1.25 g of Coomassie blue in 250 ml methanol, 50 ml acetic acid and 200 ml dH₂O. Store at room temperature.

Destaining solution (1.7 L, 10% acetic acid solution)

Add 175 ml of acetic acid and 500 ml methanol to 1.025 L of dH₂O. Store at room temperature.

Tris-acetate-EDTA (TAE) buffer 50X (1 L)

Dissolve 242 g Tris in 100 ml of 0.5 M EDTA pH 8.0 with glacial acetic acid, make up to 1 L with sterile dH₂O. Store at room temperature. Dilute to 1 X for working concentration.

Tris-borate-EDTA (TBE) buffer 5X (1 L)

Dissolve 54g of Tris base, 27.5g of Boric acid in 700 ml of dH₂O and add 20ml of 0.5M EDTA (pH 8.0). Make up to 1 L with sterile dH₂O. Dilute to 1 X for working concentration.

10 X Annealing Buffer (10ml)

Mix 1ml of Tris-HCl (100mM; pH 8.0), 1ml of NaCl (2M) and 8ml of ddH₂O. Store at room temperature.

DNA loading buffer

0.25% bromophenol blue, 0.25% xylene cyanol FF, 15% Ficoll (type 400; Pharmacia) in dH₂O. Store at room temperature.

Lysogeny Broth (LB)

Dissolve 10g Bacto-tryptone, 5g yeast extract and 10g NaCl in 800 ml of dH₂O. Adjust pH to 7.5 with NaOH. Adjust volume up to 1L. Sterilise by autoclaving and store at room temperature.

LB agar-plates

Add 15g/L of agar to LB medium and autoclave. Cool to approx. 55°C, add antibiotic and pour into petri dishes. Let harden, then invert and store at +4°C in the dark.

Terrific Broth (TB)

Dissolve 15g of Bacto-tryptone, 24g of yeast extract, 4ml of glycerol in 500ml of ddH₂O. Adjust ddH₂O up to 900 ml. Sterilise by autoclaving. Cool down to approx. 60°C and then add sterile solution of 0.17M KH₂PO₄, 0.72M K₂HPO₄ (made by dissolving 2.31 g of KH₂PO₄ and 12.54 g of K₂HPO₄ in 90 ml of ddH₂O, adjusting the volume of the solution to 100 ml with ddH₂O and sterilizing by autoclaving).

SOC Media Recipe (100 ml)

Dissolve 2g of Bacto-tryptone, 0.5g of yeast extract, 0.2ml of 5M NaCl, 0.25ml of 1M KCl, 1ml of 1M MgCl₂, 1ml of 1M MgSO₄, 2ml of 1M glucose in 90ml of ddH₂O. Adjust ddH₂O up to 100 ml. Sterilise by autoclaving.

Brilliant SYBR® Green QPCR Master Mix, Cat no. 600548

Kit Components (100-rxn kit)

2 x Brilliant SYBR Green QPCR master mix - 2 x 2.5 ml

Reference dye - 1 mM 100 µl

GST Buffer (50ml)

Mix 1.25ml of 1M Tris-HCl pH 8.0, 1.5 ml of 5M NaCl and 260µl of 1M DTT. Adjust pH to 8.0 and add ddH₂O up to 50ml. Store at 4°C.

1M Dithiothreitol (DTT)

Dissolve 1.54g in 10ml of ddH₂O.

Glutathione elution buffer (10mM glutathione (reduced) in 50mM Tris-HCl pH 8.0)

Mix 1.25ml of 1M Tris-HCl pH 8.0, 77g reduced glutathione and add ddH₂O up to 25ml. Store at -20°C.

4% non-denaturing TBA gel for EMSA (15ml)

Mix 2ml of 30% acrylamide, 1.5ml 5 X TBE buffer, 11.35ml of ddH₂O, 150µl of Ammonium persulphate and 15µl of TEMED.

1.2. Semi-quantitative RT-PCR reaction components and cycling conditions

SuperScript™ One-Step Semi-quantitative RT-PCR System Platinum® Taq, cat no. 12574-026; Kit Components (100-rxn kit)

RT/ Platinum® Taq Mix - 100 µl

2X Reaction Mix (a buffer containing 0.4 mM of each dNTP, 2.4 mM MgSO₄) - 1 ml

5 mM Magnesium Sulfate - 500 µl

5 mM Magnesium Sulfate - 1 ml

Standard reaction mix used in the RT-PCR (total volume 25 μ l)

1. 2x Reaction Mix - 12.5 μ l
2. Forward primer (10 μ M) - 0.5 μ l
3. Reverse primer (10 μ M) - 0.5 μ l
4. Template RNA (20 ng) - 1 μ l
5. RT/ Platinum[®] Taq Mix - 0.5 μ l
6. Autoclaved ddH₂O - up to 25 μ l

Thermal cycling conditions of semi-quantitative RT-PCR

	Number of cycles	Temperature and time
cDNA synthesis and pre-denaturation	1 cycle	50°C for 30min 94°C for 2min
PCR amplification	35 cycles	Denaturation: 94°C for 15s Annealing: X°C for 30s (where X is experimentally determined annealing temperature (generally T _m of primers minus 3°C)) Extension: 72°C for 1min:
Final extension	1 cycle	72°C for 10min

1.3. cDNA synthesis reaction components and thermal cycling conditions

AffinityScript Multi temperature cDNA Synthesis Kit; 200436, Agilent; Kit components:

- AffinityScript Multiple Temperature Reverse Transcriptase
- 10x AffinityScript RT buffer
- RNase Block Ribonuclease Inhibitor (40 U/ μ l)
- Oligo(dT) primer (0.5 μ g/ μ l)
- Random primers (0.1 μ g/ μ l)
- 100 mM dNTPs (25 mM each dNTP)
- RNase-free water

cDNA synthesis reaction components and thermal cycling conditions:

- Add the following to a microcentrifuge tube:
 - 1ng-5 µg of total RNA
 - RNase-free water to a total volume 15.7 µl
 - 1.0µl of oligo(dT) primer (0.5 µg/µl)
- Incubate the reaction at 65°C for 5 minutes and then cool it at room temperature (~10 minutes) to allow the primers to anneal to the RNA.
- Add the following components to a final reaction volume of 20ul
 - 2.0 µl of 10× AffinityScript RT Buffer
 - 0.8 µl of dNTP mix (25 mM each dNTP)
 - 0.5 µl of RNase Block Ribonuclease Inhibitor (40 U/µl)
 - 1 µl of AffinityScript Multiple Temperature RT
- Incubate the reactions at 42°C for 5 minutes, then at 55°C for 55min.
- Terminate the reaction by incubating at 70°C for 15 minutes.

1.4. Table showing details of primers used for semi-quantitative RT-PCR.

Name of the primer/ Gene ID	Forward/ Reverse	Primer sequence (5'→3')	length (bp)	Product length (bp)	T _m (°C)
TA05870	F	AGCGTGAATTCCTCAACTC	19	222	54.5
	R	TCAGTTCATCACCGTCAAG	19		54.5
TA14665	F	CAACCTACAGGCATTATCGC	20	229	57.3
	R	ATTGGCCATAGAATCGTCTGA	20		55.3
TA15445	F	TGGTTGTTCTGCTGGAGTTA	20	278	55.3
	R	CGTGAAGTGTGGCCAAATAA	20		55.3
TA14285	F	TACTTACCTTCGTGCCAACA	20	310	55.3
	R	TGGAGGTTGTGATTCAGGAG	20		57.3
TA16685	F	CCCACACCACATACTCTTGA	20	221	57.3
	R	ATGTATGCTGAGGGAAGTCC	20		57.3
TA21080	F	CTCAGGGTCCTATGGTCATG	20	266	59.4
	R	CAGGTGACCTTACCAATTGC	20		57.3
TA19865	F	CCTGGAACAAGAGTAGAC	18	249	53.7
	R	CCAATCACAAGGATTCGG	18		53.7
TA07475	F	AAAGAGGCTCACAAGGAACA	20	232	55.3
	R	CATGCCGTTGTGAGTAACAA	20		55.3
TA10735	F	TCTGACCTCGTCAAATGTG	19	296	54.5
	R	CGTGGGTCGTATGGAATAG	19		56.7
TA20095	F	ATGATCCATCTCGGTCCAAG	20	260	57.3
	R	TGGACCTTGACCCTTAGAGA	20		57.3
TA15705	F	TGGAGATGGAGATAGCATGC	20	239	57.3
	R	CTGGACCTCCAGATGCAC	18		58.2
TA13395	F	AGTTAACGAGCAGTGGGAAG	20	178	57.5
	R	AGCCTTATTGGTATCACTGG	20		54.5
TA10720	F	ACAATAGCAGAATCAGGAACAG	22	354	56.5
	R	TATTGGGAAACGGATGAATTCTG	23		57.1
TA13515	F	CATCCAATCCACTGTGTA	18	178	51.0
	R	TACGCTCTACTGGAGTG	17		51.8
TA11145	F	CGTTGAGGGATCTTGTGAC	19	236	56.7
	R	CTTCACACTCCTGTTCCCA	19		56.7
TA12015	F	TGTGGCCGTATTCTCGTTCC	20	318	60.0
	R	TGAAACGAGAGGGTAGGGGA	20		59.5
TA16485	F	CAAATTACTACCGAAGATTAGAG	23	169	52.5
	R	CTGTGAGTTTCTGAAAGCG	19		54.4

1.5. SYBR Green qRT-PCR reaction details and thermal cycling parameters

QRT-PCR reaction:

- 500ng cDNA
- 1µl Forward primer (0.5pM)
- 1µl Reverse Primer (0.5pM)
- 10µl SYBR Green I Master Mix
- Nuclease-free PCR-grade water to the final volume of 20µl

	Number of cycles	Temperature and time
enzyme activation and initial denaturation	1 cycle	10min at 95 °C
PCR amplification	40 cycles	Denaturation: 30 sec at 95 °C Annealing: 60 sec of at 60 °C Elongation: 60 sec at 72 °C
dissociation curve	1 cycle	60 sec at 95 °C 30 sec at 55 °C 30 sec at 95 °C

1.6. PCR conditions with Pfu polymerase and reaction components

PCR reaction (25 µl total volume)

Pfu - 0.5µl (2.5U/1µl)

dNTP - 2,5µl (2mM stock)

DNA - 2,5µl

buffer - 2,5µl (max 1/10 of total volume)

Forward primer - 1µl

Reverse primer - 1µl

ddH₂O - 15µl

PCR	Number of cycles	Temperature and time
Heated lid	1 cycle	105 °C
Initial denaturation	1 cycle	95 °C 3min
	35 cycles	95 °C 30s X °C 30s 72 °C 4min
Final extension	1 cycle	72 °C 10min
Final hold	1 cycle	10 °C

1.7. Table showing details of primers used for quantitative real time RT-PCR.

Name of the primer/ Gene ID	Forward/ Reverse	Primer sequence (5'->3')	length (bp)	Product length (bp)	Tm (°C)
TA13515	F	CGGGGAAGAGTGTA AAAAATGAGT	24		61.0
	R	G GGAGGTGATGGTCGTGATGG	20		61.4
TA11145	F	CGTTGAGGGATCTTGTGAC	19		56.7
	R	CTTCACACTCCTGTTCCCA	19		56.7
TA07100	F	GCCACCCAGTAGACCTTCA	19		58.8
	R	GTCGAGCATCAGCAAGTGT	19		56.7
TA10735	F	TCTGACCTCGTCAAATGTG	19		54.5
	R	CGTGGGTCGTATGGAATAG	19		56.7
TA15705	F	TGGAGATGGAGATAGCATGC	20		57.3
	R	CTGGACCTCCAGATGCAC	18		58.2
TA11610	F	ACGCAAATGGAATCCTCAAC	20		55.3
	R	TATTCGTCGTGCTCTGCTAA	20		55.3
TA10720	F	ACAATAGCAGAATCAGGAACAG	22		56.5
	R	TATTGGGAAACGGATGAATTCTG	23		57.1

1.8. Table showing primers with restriction sites used for cloning.

Primer name	Sequence 5'->3'	Length (bp)	Tm (°C)
TA13515D- EcorRI_F	CAGGAATTCGTACAGGGTATGGTTGGATATTCT	33	58.8
TA13515D- XhoI_R	GCACTCGAGGCTGAATACGCTCTACTGGAGTGC	33	73.2
TA12015D- EcoRI_F	CAGGAATTCTACCGAAGGAAGCCAATCTCATC	32	68.2
TA12015D- XhoI_R	GCACTCGAGAGATGTGGTTCCTCTCGGT	29	69.5
TA11145D- EcorRI_F	CAGGAATTCCAAAGAACGAGCGCAAAGATTC	32	66.9
TA11145D- XhoI_R	GTTCTCGAGTGTTAAATCTTATCATTATGTCTAAGTGC	38	66.2
TA16485D- EcorRI_F	CAGGAATTCAGAGCAAATTACTACCGAAGATTAG	34	65.9
TA16485D- XhoI_R	GCACTCGAGCGGTCAGAT TGTGGTTGGT TTC TG	35	71.8

1.9. Table showing list of oligos used for EMSA experiment.

Oligo name	Sequence 5'→3'	Description
Bio-2x(A)CACAC(A)_F	Bio- GATACACACTTATGCACACACA	Biotynylated target motif for TA11145; Upstream region of TA11145; forward strand
2x(A)CACAC(A)_F	GATACACACTTATGCACACACA	Target motif for TA11145; Upstream region of TA11145
2x(A)CACAC(A)_R	TGTGTGTGCATAAGTGTGTATC	Target motif for TA11145; Upstream region of TA11145, reverse strand
(A)CACAC(A)_MUT_F	Bio - GATACACACTTATGCAGAATAT	Biotynylated mutated target motif for TA11145; upstream region of TA11145; forward strand
(A)CACAC(A)_MUT_F	GATACACACTTATGCAGAATAT	Mutated target motif for TA11145; upstream region of TA11145; reverse strand
(A)CACAC(A)_MUT_R	ATATTCTGCATAAGTGTGTATC	Mutated biotinylated target motif for TA11145; upstream region of TA11145; reverse primer
Bio(A)CACAC(A)- 2xMUT_F	Bio - GATATAGAATTATGCAGAATAT	Double mutated biotinylated target motif for TA11145; upstream region of TA11145; forward primer
(A)CACAC(A)- 2xMUT_F	GATATAGAATTATGCAGAATAT	Double mutated target motif for TA11145; upstream region of TA11145; forward primer
(A)CACAC(A)- 2MUT_R:	ATATTCTGCATAATTCTATATC	Double mutated target motif for TA11145; upstream region of TA11145; reverse strand
Bio-GTGTAC_F	Bio- AATATTATAATAGTCGTAGCCAT CAATGTGTACACATGGTAATAT AGATTTTCGTTTATATT	Biotynylated target motif of PBANKA_143750/PFL1085w/TA13515, forward primer
Bio-NoMotif_F	Bio- AATATTATAATAGTCGTAGCCAT CAATAAAAAAAAAATGGTAATATA GATTTTCGTTTATATT	Biotynylated no motif oligo (negative control), forward primer
Bio-GTGTAC-MUT_F	Bio- AATATTATAATAGTCGTAGCCAT CAATATATAAAAAATGGTAATATA GATTTTCGTTTATATT	Biotynylated mutated motif of PBANKA_143750/PFL1085w/TA13515 (negative control; G/C replaced with A); forward primer
Bio-TGCATGCA_F	Bio- AATATTATAATAGTCGTAGCCAT CAATTGCATGCAATGGTAATAT AGATTTTCGTTTATATT	Biotynylated another ApiAP2 motif (negative control; PBANKA_132980/PF14_0633), forward primer
GTGTAC_F	AATATTATAATAGTCGTAGCCAT CAATGTGTACACATGGTAATAT AGATTTTCGTTTATATT	Target motif (PBANKA_143750/PFL1085w/TA13515; forward primer)
NoMotif_F	AATATTATAATAGTCGTAGCCAT CAATAAAAAAAAAATGGTAATATA GATTTTCGTTTATATT	No motif oligo (negative control), forward primer
GTGTAC-MUT_F	AATATTATAATAGTCGTAGCCAT CAATATATAAAAAATGGTAATATA GATTTTCGTTTATATT	Mutated motif (negative control; G/C replaced with A)
TGCATGCA_F	AATATTATAATAGTCGTAGCCAT CAATTGCATGCAATGGTAATAT AGATTTTCGTTTATATT	Another ApiAP2 motif (negative control; PBANKA_132980/PF14_0633)
GTGTAC_R	AATATAAACGAAAATCTATATTA CCATGTGTACACATTGATGGCT ACGACTATTATAATATT	REVERSE Target motif (PBANKA_143750/PFL1085w/TA13515)
NoMotif_R	AATATAAACGAAAATCTATATTA CCATTTTTTTTTTATTGATGGCTA	REVERSE No motif (negative control)

	CGACTATTATAATATT	
GTGTAC-MUT_R	AATATAAACGAAAATCTATATTA CCATTTTTATATATTGATGGCTA CGACTATTATAATATT	REVERSE mutated motif (negative control; G/C replaced with A)
TGCATGCA_R	AATATAAACGAAAATCTATATTA CCATTGCATGCAATTGATGGCT ACGACTATTATAATATT	REVERSE Another ApiAP2 motif (negative control; PBANKA_132980/PF14_0633)
Bio- TA13515Target_F	Bio- ATGAGATAAATGTGTACGTAGG AATCTTCA	Biotinylated target motif for TA13515 ; TA10735 (GATA TF) upstream region; forward primer
TA13515Target_F	ATGAGATAAATGTGTACGTAGG AATCTTCA	Target motif for TA13515 ; TA10735 (GATA TF) upstream region, forward primer
TA13515Target_R	TGAAGATTCCTACGTACACATTT ATCTCAT	Reverse target motif for TA13515, TA10735 (GATA TF) upstream region
Bio-TA11145D_F	BIO- AATATTATAATAGTCGTAGCCAT CAATCCACACACCATGGTAATAT AGATTTTCGTTTATATT	Biotinylated target motif for TA11145, forward primer
TA11145D_F	AATATTATAATAGTCGTAGCCAT CAATCCACACACCATGGTAATAT AGATTTTCGTTTATATT	Target motif for TA11145, forward primer
TA11145D_R	AATATAAACGAAAATCTATATTA CCATGGTGTGTGGATTGATGGC TACGACTATTATAATATT	Reverse target motif for TA11145
Bio-Tash1u_F	BIO- ATTGTTAATTCATCCAGATC TATAAAA	Biotinylated oligo containing C-box and TCTATA motif; upstream region of TA03125; forward primer
Tash1u_F	ATTGTTAATTCATCCAGATC TATAAAA	Oligo containing C-box and TCTATA motif; upstream region of TA03125; forward primer
Tash1u_R	TTTTATAGATCTGGATGGGGAA TTAACAAT	Reverse oligo containing C-box and TCTATA motif; upstream region of TA03125

Appendix 2 - Rank Product results from Chapter 2

2.1. Full list of genes up-regulated from macroschizont to merozoite obtained by Rank Product (n=152, FDR≤0.05)

Gene ID	Annotation	FC	RP score	EE	FDR
TA05870	rhoptry-associated protein, putative	5,12	3,13E-31	0	0
TA14665	hypothetical protein	5,133333	1,41E-30	0	0
TA08360	hypothetical protein, conserved	4,636667	1,69E-28	0	0
TA21080	Map2 kinase, putative	4,156667	2,40E-26	0	0
TA05340	hypothetical protein, conserved	4,033333	1,21E-25	0	0
TA16660	hypothetical protein, conserved	3,956667	2,36E-25	0	0
TA05495	hypothetical protein	3,893333	2,65E-25	0	0
TA13045	hypothetical protein, conserved	3,673333	5,49E-24	0	0
TA13825	hypothetical protein	3,63	1,28E-23	0	0
TA07585	hypothetical protein	3,506667	2,58E-23	0	0
TA19390	hypothetical protein, conserved	3,396667	1,11E-22	0	0
TA18005	hypothetical protein	3,383333	2,06E-22	0	0
TA19040	hypothetical protein, conserved	3,353333	2,30E-22	0	0
TA19445	hypothetical protein, conserved	3,243333	8,27E-22	0	0
TA11905	hypothetical protein	3,426667	9,64E-22	0	0
TA20020	hypothetical protein, conserved	3,076667	4,25E-21	0,02	0,00125
TA14680	hypothetical protein	3,016667	9,70E-21	0,03	0,00176
TA21400	hypothetical protein	2,926667	2,34E-20	0,04	0,00222
TA17325	integral membrane protein, putative	2,953333	2,41E-20	0,04	0,00211
TA16375	hypothetical protein, conserved	2,943333	3,47E-20	0,04	0,002
TA05760	rhoptry-associated protein, putative	2,906667	3,70E-20	0,05	0,00238
TA21395	hypothetical protein	2,946667	4,08E-20	0,05	0,00227
TA04105	cysteine proteinase, putative	2,896667	5,75E-20	0,07	0,00304
TA13515	hypothetical protein, conserved	2,88	8,26E-20	0,07	0,00292
TA14955	hypothetical protein	2,846667	1,43E-19	0,08	0,0032
TA13215	hypothetical protein, conserved	2,79	1,43E-19	0,08	0,00308
TA16485	hypothetical protein, conserved	2,776667	2,13E-19	0,09	0,00333
TA11455	hypothetical protein, conserved	2,756667	2,77E-19	0,09	0,00321
TA18855	Sfil-subtelomeric fragment related protein family member, putative	2,79	3,15E-19	0,1	0,00345
TA16420	hypothetical protein	2,833333	3,28E-19	0,1	0,00333
TA18195	hypothetical protein	2,753333	4,13E-19	0,12	0,00387
TA15485	hypothetical protein, conserved	2,743333	4,60E-19	0,13	0,00406
TA14205	hypothetical protein	2,67	9,74E-19	0,16	0,00485
TA16155	hypothetical protein	2,66	1,04E-18	0,16	0,00471
TA17100	hypothetical protein	2,673333	1,20E-18	0,17	0,00486
TA04660	hypothetical protein, conserved	2,64	1,28E-18	0,17	0,00472
TA21390	Theileria parva Tpr-related protein, putative	2,653333	1,29E-18	0,17	0,00459
TA11285	hypothetical protein	2,7	1,62E-18	0,19	0,005
TA17490	hypothetical protein	2,653333	2,07E-18	0,19	0,00487
TA08480	hypothetical protein, conserved	2,613333	2,91E-18	0,21	0,00525
TA07985	hypothetical protein	2,61	4,77E-18	0,22	0,00537
TA07025	Tpr-related protein family member, putative	2,56	5,58E-18	0,23	0,00548
TA21385	hypothetical protein	2,536667	5,62E-18	0,23	0,00535
TA03260	hypothetical protein	2,53	6,10E-18	0,23	0,00523
TA20555	myosin a, putative	2,493333	1,09E-17	0,3	0,00667
TA19610	hypothetical protein, conserved	2,49	1,91E-17	0,34	0,00739
TA14210	hypothetical protein	2,443333	1,92E-17	0,34	0,00723
TA07920	hypothetical protein	2,456667	2,96E-17	0,43	0,00896
TA15445	Tpr-related protein family member,	2,403333	3,39E-17	0,48	0,0098

	putative				
TA15620	hypothetical protein	2,396667	4,06E-17	0,52	0,0104
TA15625	hypothetical protein	2,413333	4,21E-17	0,52	0,0102
TA19275	hypothetical protein	2,42	4,43E-17	0,53	0,01019
TA14285	Sfil-subtelomeric fragment related protein family member, putative	2,413333	4,53E-17	0,53	0,01
TA07630	hypothetical protein, conserved	2,393333	5,90E-17	0,56	0,01037
TA16685	polymorphic antigen precursor-like protein, putative	2,393333	6,37E-17	0,56	0,01018
TA13890	hypothetical protein	2,32	2,07E-16	0,83	0,01482
TA04565	hypothetical protein	2,303333	2,20E-16	0,84	0,01474
TA19505	hypothetical protein, conserved	2,303333	2,37E-16	0,84	0,01448
TA17055	hypothetical protein	2,29	2,42E-16	0,84	0,01424
TA08235	hypothetical protein	2,303333	2,42E-16	0,84	0,014
TA14310	hypothetical protein	2,276667	3,29E-16	1	0,01639
TA13540	hypothetical protein	2,27	3,34E-16	1	0,01613
TA17358		2,253333	3,82E-16	1,01	0,01603
TA13530	phosphate transporter, putative	2,246667	3,97E-16	1,01	0,01578
TA15490	hypothetical protein	2,253333	4,84E-16	1,04	0,016
TA17115	Sfil-subtelomeric fragment related protein family member, putative	2,25	5,20E-16	1,07	0,01621
TA20435	hypothetical protein	2,233333	5,44E-16	1,12	0,01672
TA06795	hypothetical protein, conserved	2,226667	5,55E-16	1,14	0,01676
TA18275	hypothetical protein	2,223333	5,91E-16	1,17	0,01696
TA20150	Theileria-specific hypothetical protein	2,22	6,00E-16	1,17	0,01671
TA14130	Tpr-related protein family member, putative	2,226667	6,31E-16	1,19	0,01676
TA04355	hypothetical protein	2,19	8,91E-16	1,31	0,01819
TA13940	hypothetical protein	2,21	9,32E-16	1,36	0,01863
TA07435	Sfil-subtelomeric fragment related protein family member, putative	2,153333	1,15E-15	1,47	0,01986
TA10645	hypothetical protein	2,176667	1,20E-15	1,49	0,01987
TA20985	hypothetical protein	2,153333	1,30E-15	1,55	0,02039
TA15095	Tpr-related protein family member, putative	2,176667	1,35E-15	1,55	0,02013
TA15090	Tpr-related protein family member, putative	2,166667	1,37E-15	1,55	0,01987
TA15355	hypothetical protein	2,15	1,51E-15	1,62	0,02051
TA19675	hypothetical protein	2,153333	1,58E-15	1,65	0,02063
TA17960	hypothetical protein	2,15	1,62E-15	1,65	0,02037
TA13535	hypothetical protein	2,106667	1,62E-15	1,65	0,02012
TA05375	hypothetical protein	2,163333	1,67E-15	1,71	0,0206
TA14335	hypothetical protein	2,15	1,72E-15	1,73	0,0206
TA10690	ubiquitin-conjugating enzyme E2, putative	2,116667	2,36E-15	1,88	0,02212
TA19975	integral membrane protein, putative	2,106667	2,73E-15	1,95	0,02267
TA03850	Tpr-related protein family member, putative	2,126667	2,85E-15	1,98	0,02276
TA14120	Tpr-related protein family member, putative	2,093333	3,18E-15	2,04	0,02318
TA03540	hypothetical protein	2,06	3,40E-15	2,12	0,02382
TA14135	Tpr-related protein family member, putative	2,103333	3,87E-15	2,15	0,02389
TA03755	sporozoite surface antigen (spag -1)	2,083333	3,87E-15	2,15	0,02363
TA17500	Sfil-subtelomeric fragment related protein family member, putative	2,093333	3,92E-15	2,16	0,02348
TA13345	hypothetical protein	2,07	5,14E-15	2,32	0,02495
TA19575	hypothetical protein, conserved	2,06	5,66E-15	2,37	0,02521
TA09600	Tpr-related protein family member, putative	2,01	7,70E-15	2,61	0,02747

TA03480	hypothetical protein, conserved	2,033333	8,04E-15	2,61	0,02719
TA19075	hypothetical protein	2,026667	8,17E-15	2,63	0,02711
TA21050	hypothetical protein	2,016667	8,20E-15	2,63	0,02684
TA03855	Tpr-related protein family member, putative	2,026667	9,80E-15	2,74	0,02768
TA12015	hypothetical protein	2,003333	1,02E-14	2,79	0,0279
TA03845	hypothetical protein, conserved	2	1,04E-14	2,8	0,02772
TA15500	Tpr-related protein family member, putative	1,996667	1,10E-14	2,84	0,02784
TA17705	hypothetical protein	1,996667	1,24E-14	2,93	0,02845
TA17595	Sfil-subtelomeric fragment related protein family member, putative	2,013333	1,26E-14	2,95	0,02837
TA14835	hypothetical protein	2	1,45E-14	3,08	0,02933
TA14125	hypothetical protein	1,983333	1,48E-14	3,09	0,02915
TA15405	hypothetical protein	1,92	1,63E-14	3,18	0,02972
TA19115	Sfil-subtelomeric fragment related protein family member, putative	1,986667	1,72E-14	3,2	0,02963
TA21345	hypothetical protein	1,943333	1,74E-14	3,25	0,02982
TA21380	hypothetical protein	1,976667	1,76E-14	3,27	0,02973
TA04370	hypothetical protein, conserved	1,973333	2,16E-14	3,5	0,03153
TA08470	cyclin-dependent serine/threonine kinase (Cdk)-related protein, putative	1,95	2,33E-14	3,55	0,0317
TA11680	hypothetical membrane protein, conserved	1,94	2,41E-14	3,58	0,03168
TA07305	hypothetical protein	1,963333	2,45E-14	3,59	0,03149
TA20220	hypothetical protein, conserved	1,94	2,47E-14	3,6	0,0313
TA21205	hypothetical protein	1,966667	2,55E-14	3,64	0,03138
TA03475	hypothetical protein	1,956667	2,59E-14	3,66	0,03128
TA15320	hypothetical protein, conserved	1,923333	2,74E-14	3,72	0,03153
TA18780	hypothetical protein	1,93	2,77E-14	3,73	0,03134
TA05615	hypothetical protein	1,92	2,78E-14	3,73	0,03108
TA03300	hypothetical protein	1,936667	2,87E-14	3,75	0,03099
TA04790	Tpr-related protein family member, putative	1,96	2,92E-14	3,76	0,03082
TA15350	hypothetical protein	1,916667	3,61E-14	4,12	0,0335
TA06765	hypothetical protein, conserved	1,916667	4,15E-14	4,36	0,03516
TA17685	aspartyl(acid) protease, putative	1,906667	4,32E-14	4,42	0,03536
TA18600	hypothetical protein	1,89	4,63E-14	4,56	0,03619
TA16050	Sfil-subtelomeric fragment related protein family member, putative	1,873333	4,78E-14	4,6	0,03622
TA15660	cysteine protease, putative	1,903333	5,18E-14	4,75	0,03711
TA09115	hypothetical protein	1,876667	5,22E-14	4,75	0,03682
TA18980	hypothetical protein	1,893333	5,32E-14	4,78	0,03677
TA15815	hypothetical protein	1,816667	5,67E-14	4,83	0,03687
TA19530	hypothetical protein	1,88	5,97E-14	4,95	0,0375
TA13665	hypothetical protein	1,846667	6,53E-14	5,11	0,03842
TA04130	hypothetical protein, conserved	1,866667	6,77E-14	5,16	0,03851
TA04635	hypothetical protein	1,866667	7,63E-14	5,36	0,0397
TA14605	Tpr-related protein family member, putative	1,85	1,01E-13	5,91	0,04346
TA08525	hypothetical protein, conserved	1,833333	1,08E-13	6,02	0,04394
TA17491		1,81	1,20E-13	6,17	0,04471
TA05705	rhoptry-associated protein, putative	1,81	1,29E-13	6,4	0,04604
TA11195	hypothetical protein	1,803333	1,34E-13	6,46	0,04614
TA18295	Theileria-specific hypothetical protein	1,823333	1,42E-13	6,55	0,04645
TA03640	hypothetical protein	1,83	1,51E-13	6,67	0,04697
TA21370	hypothetical protein	1,77	1,51E-13	6,67	0,04664
TA07955	hypothetical protein	1,79	1,69E-13	6,89	0,04785
TA07162		1,786667	1,92E-13	7,16	0,04938

TA14140	Tpr-related protein family member, putative	1,783333	2,05E-13	7,3	0,05
TA05245	Tpr-related protein family member, putative	1,776667	2,07E-13	7,32	0,0498
TA03521	Theileria-specific hypothetical telomeric sfi fragment-related protein	1,78	2,12E-13	7,34	0,04959
TA21375	hypothetical protein	1,79	2,15E-13	7,4	0,04966
TA04825	hypothetical protein, conserved	1,773333	2,29E-13	7,51	0,05007
TA04795	hypothetical protein	1,786667	2,32E-13	7,55	0,05
TA12440	hypothetical protein	1,773333	2,33E-13	7,57	0,0498

2.2. Full list of genes down-regulated from macroschizont to merozoite obtained by Rank Product (n=115, FDR≤0.05)

Gene ID	Annotation	FC	RP score	EE	FDR
TA19865	surface protein d precursor	- 3,383333333	7,04E-31	0	0
TA11410	Theileria-specific sub-telomeric protein, SVSP family, putative	- 3,203333333	1,01E-28	0	0
TA11405	subtelomeric sfi-fragment-related protein family member, putative	- 3,116666667	6,57E-28	0	0
TA15705	hypothetical protein (ta9)	-3,1	2,17E-27	0	0
TA09805	Theileria-specific sub-telomeric protein, SVSP family	-2,93	7,76E-26	0	0
TA10505	hypothetical protein	- 2,913333333	1,76E-25	0	0
TA09790	Theileria-specific sub-telomeric protein, SVSP family	-2,69	8,78E-24	0	0
TA18010	integral membrane protein, putative	- 2,693333333	1,44E-23	0	0
TA09420	Theileria-specific sub-telomeric protein, SVSP family, putative	- 2,613333333	2,04E-23	0	0
TA02480	hexose transporter (HT1 homologue), putative	- 2,613333333	3,93E-23	0	0
TA05580	Theileria-specific sub-telomeric protein, SVSP family	-2,59	4,48E-23	0	0
TA15695	hypothetical protein	- 2,636666667	4,76E-23	0	0
TA09810	Theileria-specific sub-telomeric protein, SVSP family	- 2,543333333	6,70E-23	0	0
TA09435	Theileria-specific sub-telomeric protein, SVSP family, putative	- 2,516666667	3,11E-22	0	0
TA15710	hypothetical protein	-2,5	4,99E-22	0	0
TA18895	conserved Theileria-specific sub-telomeric protein, SVSP family	- 2,433333333	7,76E-22	0	0
TA09430	Theileria-specific sub-telomeric protein, SVSP family, putative	-2,4	1,95E-21	0,01	0,000588235
TA17555	Theileria-specific sub-telomeric protein, SVSP family	- 2,396666667	2,15E-21	0,01	0,000555556
TA11940	hypothetical protein	- 2,386666667	2,68E-21	0,01	0,000526316
TA09815	Sfil-subtelomeric fragment related protein family member, putative	- 2,363333333	3,82E-21	0,02	0,001
TA09800	Theileria-specific sub-telomeric protein, SVSP family	- 2,373333333	4,30E-21	0,02	0,000952381
TA10530	hypothetical protein	- 2,353333333	5,18E-21	0,02	0,000909091
TA17545	Theileria-specific sub-telomeric protein, SVSP family	-2,25	4,52E-20	0,05	0,002173913

TA05575	Theileria-specific sub-telomeric protein, SVSP family	-	5,32E-20	0,05	0,002083333
	2,243333333				
TA19005	conserved Theileria-specific sub-telomeric protein, SVSP family	-2,23	7,05E-20	0,07	0,0028
TA17125	Theileria-specific sub-telomeric protein, SVSP family	-	1,21E-19	0,08	0,003076923
	2,193333333				
TA09505	Sfil-subtelomeric fragment related protein family member, putative	-	2,01E-19	0,09	0,003333333
	2,183333333				
TA18890	conserved Theileria-specific sub-telomeric protein, SVSP family	-2,17	2,72E-19	0,09	0,003214286
TA20095	Tashat2 protein	-	6,55E-19	0,15	0,005172414
	2,126666667				
TA02905	hypothetical protein	-	8,61E-19	0,16	0,005333333
	2,043333333				
TA16035	Theileria-specific sub-telomeric protein, SVSP family, putative	-	8,76E-19	0,16	0,00516129
	2,086666667				
TA16030	Theileria-specific sub-telomeric protein, SVSP family, putative	-2,07	1,10E-18	0,17	0,0053125
TA09510	Sfil-subtelomeric fragment related protein family member, putative	-	1,21E-18	0,17	0,005151515
	1,976666667				
TA12265	Sfil-subtelomeric fragment related protein family member, putative	-2,08	1,23E-18	0,17	0,005
TA11945	hypothetical protein	-	1,73E-18	0,19	0,005428571
	2,056666667				
TA04675	hypothetical protein	-1,98	1,93E-18	0,19	0,005277778
TA11400	Sfil-subtelomeric fragment related protein family member, putative	-2,05	2,20E-18	0,2	0,005405405
TA16090	glutenin, putative	-	2,95E-18	0,21	0,005526316
	1,966666667				
TA17485	hypothetical protein	-	4,18E-18	0,21	0,005384615
	1,903333333				
TA05540	Theileria-specific sub-telomeric protein, SVSP family	-	5,41E-18	0,23	0,00575
	2,013333333				
TA10420	hypothetical protein	-	5,90E-18	0,23	0,005609756
	1,926666667				
TA05960	Tpr-related protein family member, putative	-1,88	7,87E-18	0,28	0,006666667
TA18950	conserved Theileria-specific sub-telomeric protein, SVSP family	-	1,15E-17	0,3	0,006976744
	1,963333333				
TA15015	Tpr-related protein family member, putative	-	1,31E-17	0,3	0,006818182
	1,943333333				
TA05565	Theileria-specific sub-telomeric protein, SVSP family	-	2,12E-17	0,37	0,008222222
	1,926666667				
TA18865	conserved Theileria-specific sub-telomeric protein, SVSP family	-1,93	2,15E-17	0,37	0,008043478
TA17915	hypothetical protein	-1,82	2,96E-17	0,43	0,009148936
TA16025	Theileria-specific sub-telomeric protein, SVSP family, putative	-	3,02E-17	0,43	0,008958333
	1,913333333				
TA07475	Sfil-subtelomeric fragment related protein family member, putative	-	3,60E-17	0,49	0,01
	1,813333333				
TA18540	hypothetical protein	-	3,99E-17	0,51	0,0102
	1,766666667				
TA05150	hypothetical protein, conserved	-	5,19E-17	0,54	0,010588235
	1,873333333				

TA17375	polymorphic antigen precursor, putative	-1,79	5,83E-17	0,56	0,010769231
TA11950	hypothetical P-, Q-rich protein family protein, putative	1,783333333	6,87E-17	0,56	0,010566038
TA16185	haloacid dehalogenase-like family hydrolase, putative	1,776666667	7,17E-17	0,59	0,010925926
TA09425	Theileria-specific sub-telomeric protein, SVSP family, putative	-1,84	9,61E-17	0,67	0,012181818
TA11385	Theileria-specific sub-telomeric protein, SVSP family, putative	-1,84	1,03E-16	0,67	0,011964286
TA17120	Theileria-specific sub-telomeric protein, SVSP family	-1,83	1,34E-16	0,71	0,01245614
TA04470	hypothetical protein	-1,72	1,68E-16	0,81	0,013965517
TA05560	Theileria-specific sub-telomeric protein, SVSP family	1,773333333	2,21E-16	0,84	0,014237288
TA19060	conserved Theileria-specific sub-telomeric protein, SVSP family	-1,72	2,22E-16	0,84	0,014
TA05965	Tpr-related protein family member, putative	1,646666667	2,22E-16	0,84	0,013770492
TA05555	Theileria-specific sub-telomeric protein, SVSP family	1,766666667	2,31E-16	0,84	0,013548387
TA10655	hypothetical protein, conserved	1,713333333	2,50E-16	0,85	0,013492063
TA17550	Theileria-specific sub-telomeric protein, SVSP family	1,783333333	2,56E-16	0,87	0,01359375
TA05570	Theileria-specific sub-telomeric protein, SVSP family	1,793333333	2,64E-16	0,89	0,013692308
TA03125	Tash1-like protein, putative	-1,69	2,89E-16	0,94	0,014242424
TA11415	Sfil-subtelomeric fragment related protein family member, putative	1,766666667	3,73E-16	1,01	0,015074627
TA17540	Theileria-specific sub-telomeric protein, SVSP family	1,763333333	3,77E-16	1,01	0,014852941
TA13350	hypothetical protein	-1,74	4,01E-16	1,01	0,014637681
TA04895	hypothetical protein, conserved	1,613333333	4,98E-16	1,05	0,015
TA15580	hypothetical protein	-1,57	5,57E-16	1,14	0,016056338
TA11935	hypothetical protein	1,746666667	5,69E-16	1,16	0,016111111
TA03120	Tash-like protein, putative	1,636666667	5,84E-16	1,17	0,016027397
TA16555	hypothetical protein	1,623333333	6,38E-16	1,19	0,016081081
TA05545	Theileria-specific sub-telomeric protein, SVSP family	1,723333333	8,30E-16	1,29	0,0172
TA13965	hypothetical protein	1,553333333	1,14E-15	1,46	0,019210526
TA09785	Theileria-specific sub-telomeric protein, SVSP family	-1,7	1,30E-15	1,54	0,02
TA17535	Theileria-specific sub-telomeric protein, SVSP family	-1,64	1,31E-15	1,55	0,019871795
TA16820	hypothetical protein	-1,57	1,54E-15	1,62	0,020506329
TA10865	Tpr-related protein family member, putative	1,546666667	1,56E-15	1,64	0,0205
TA03535	hypothetical protein	1,633333333	1,61E-15	1,65	0,02037037
TA09865	Theileria-specific sub-telomeric protein, SVSP family	1,596666667	2,04E-15	1,77	0,021585366
TA12370	hypothetical protein, conserved	1,516666667	2,49E-15	1,92	0,02313253
TA17346	hypothetical protein	-1,48	3,44E-15	2,12	0,025238095
TA04550	hypothetical protein	-1,49	3,81E-15	2,14	0,025176471

TA10735	GATA-specific transcription factor, putative	-	4,85E-15	2,28	0,026511628
		1,496666667			
TA10755	hypothetical protein, conserved	-	5,10E-15	2,32	0,026666667
		1,473333333			
TA09025	hypothetical protein, conserved	-	5,73E-15	2,37	0,026931818
		1,446666667			
TA03145	Tash1-like protein, putative	-	5,91E-15	2,39	0,026853933
		1,526666667			
TA08365	hypothetical protein	-1,44	7,29E-15	2,54	0,028222222
TA12635	Theileria-specific hypothetical protein family member, putative	-1,43	7,89E-15	2,61	0,028681319
TA09795	Theileria-specific sub-telomeric protein, SVSP family	-	9,30E-15	2,72	0,029565217
		1,563333333			
TA11390	Theileria-specific sub-telomeric protein, SVSP family, putative	-	1,43E-14	3,07	0,033010753
		1,466666667			
TA11960	hypothetical P-,Q-rich family protein, putative	-	1,88E-14	3,34	0,035531915
		1,406666667			
TA10770	hypothetical protein, conserved	-1,36	2,18E-14	3,51	0,036947368
TA12945	hypothetical protein, conserved	-	2,32E-14	3,54	0,036875
		1,393333333			
TA12525	Theileria-specific hypothetical protein family member, putative	-	2,87E-14	3,75	0,038659794
		1,353333333			
TA06330	Tpr-related protein family member, putative	-	3,54E-14	4,08	0,041632653
		1,353333333			
TA06060	hypothetical protein, conserved	-	3,79E-14	4,26	0,043030303
		1,346666667			
TA10800	hypothetical protein, conserved	-	4,20E-14	4,39	0,0439
		1,346666667			
TA11605	hypothetical protein	-	4,23E-14	4,41	0,043663366
		1,313333333			
TA12315	hypothetical protein	-	4,83E-14	4,63	0,045392157
		1,343333333			
TA06355	hypothetical protein	-	5,05E-14	4,71	0,045728155
		1,333333333			
TA21060	hypothetical protein	-	5,66E-14	4,83	0,046442308
		1,326666667			
TA20990	hypothetical protein	-	5,73E-14	4,86	0,046285714
		1,356666667			
TA13925	hypothetical protein	-	7,07E-14	5,22	0,049245283
		1,286666667			
TA12580	Theileria-specific hypothetical protein family member, putative	-1,3	7,43E-14	5,33	0,049813084
TA04620	hypothetical protein, conserved	-	7,63E-14	5,36	0,04962963
		1,313333333			
TA09360	hypothetical protein, conserved	-	7,84E-14	5,43	0,049816514
		1,283333333			
TA03165	Tash1-like protein, putative	-	8,03E-14	5,46	0,049636364
		1,383333333			
TA07725	hypothetical protein	-	8,06E-14	5,46	0,049189189
		1,236666667			
TA17135	Theileria-specific sub-telomeric protein, SVSP family	-	8,47E-14	5,53	0,049375
		1,286666667			
TA20460	tryptophanyl-trna synthetase, putative	-1,29	8,60E-14	5,57	0,049292035
TA18335	hypothetical protein	-	8,74E-14	5,6	0,049122807
		1,323333333			
TA06360	Tpr-related protein family member, putative	-	8,92E-14	5,63	0,048956522
		1,346666667			

2.3. Full list of genes up-regulated from sporozoite to macroschizont obtained by Rank Product (n=67; FDR≤0.05)

Gene ID	Annotation	FC	RP score	EE	FDR
TA16900	Tpr-related protein family member, putative	5,846666667	1,49E-31	0	0
TA08535	Tpr-related protein family member, putative	5,44	1,12E-29	0	0
TA08540	Tpr-related protein family member, putative	5,56	1,20E-29	0	0
TA16895	Tpr-related protein family member, putative	5,39	3,81E-28	0	0
TA16870	Tpr-related protein family member, putative	4,786666667	5,06E-26	0	0
TA19865	surface protein d precursor	4,503333333	8,25E-26	0	0
TA16875	Tpr-related protein family member, putative	4,116666667	1,26E-24	0	0
TA15705	hypothetical protein	3,75	1,32E-23	0	0
TA06635	Tpr-related protein family member, putative	3,37	1,48E-23	0	0
TA06580	Tpr-related protein family member, putative	3,236666667	4,97E-23	0,01	0,001
TA16765	Tpr-related protein family member, putative	3,233333333	6,92E-23	0,01	0,000909091
TA06260	Tpr-related protein family member, putative	3	8,45E-22	0,02	0,001666667
TA10420	hypothetical protein	3,606666667	8,97E-22	0,02	0,001538462
TA16790	Tpr-related protein family member, putative	2,79	2,12E-21	0,02	0,001428571
TA06880	Tpr-related protein family member, putative	2,663333333	6,22E-21	0,04	0,002666667
TA06330	Tpr-related protein family member, putative	2,63	7,45E-21	0,04	0,0025
TA05605	hypothetical protein, conserved	3,03	1,19E-20	0,04	0,002352941
TA06740	Tpr-related protein family member, putative	2,62	1,67E-20	0,04	0,002222222
TA06675	hypothetical protein	2,52	3,75E-20	0,05	0,002631579
TA08365	hypothetical protein	2,5	4,35E-20	0,06	0,003
TA19824	hypothetical protein	2,473333333	9,67E-20	0,08	0,003809524
TA05965	Tpr-related protein family member, putative	2,436666667	2,72E-19	0,11	0,005
TA06895	Tpr-related protein family member, putative	2,4	2,79E-19	0,11	0,004782609
TA02905	hypothetical protein	2,39	2,97E-19	0,12	0,005
TA06265	Tpr-related protein family member, putative	2,28	2,85E-18	0,23	0,0092
TA04235	Tpr-related protein family member, putative	2,19	4,33E-18	0,24	0,009230769
TA06745	Tpr-related protein family member, putative	2,23	5,81E-18	0,28	0,01037037
TA04240	Tpr-related protein family member, putative	2,186666667	6,27E-18	0,28	0,01
TA15695	hypothetical protein	2,61	6,91E-18	0,32	0,011034483
TA06900	Tpr-related protein family	2,19	9,48E-	0,36	0,012

	member, putative		18		
TA06925	hypothetical protein, conserved	2,18	9,74E-18	0,38	0,012258065
TA07065	chaperonin (HSP60), putative	2,106666667	1,10E-17	0,38	0,011875
TA05960	Tpr-related protein family member, putative	2,233333333	1,48E-17	0,4	0,012121212
TA15580	hypothetical protein	2,23	2,62E-17	0,45	0,013235294
TA08775	bifunctional dihydrofolate reductase/thymidilate synthase, putative	2,003333333	6,34E-17	0,57	0,016285714
TA10530	hypothetical protein	1,983333333	8,03E-17	0,61	0,016944444
TA06535	hypothetical protein	2,643333333	8,82E-17	0,62	0,016756757
TA16795	Tpr-related protein family member, putative	2,03	9,33E-17	0,63	0,016578947
TA09790	Theileria-specific sub-telomeric protein, SVSP family	2,146666667	1,83E-16	0,73	0,018717949
TA08675	hypothetical protein, conserved	1,966666667	2,08E-16	0,73	0,01825
TA13085	hypothetical protein	2,056666667	3,72E-16	0,89	0,021707317
TA10475	hypothetical protein	2,096666667	6,42E-16	1,04	0,024761905
TA09340	hypothetical protein, conserved	1,99	6,48E-16	1,04	0,024186047
TA13090	hypothetical protein	2,02	6,70E-16	1,06	0,024090909
TA13970	hypothetical protein, conserved	1,916666667	8,71E-16	1,14	0,025333333
TA20945	ATP synthase beta chain, mitochondrial precursor, putative	1,86	1,08E-15	1,27	0,027608696
TA06400	hypothetical protein, conserved	1,846666667	1,23E-15	1,35	0,028723404
TA16910	hypothetical protein	1,873333333	1,26E-15	1,35	0,028125
TA13330	hypothetical protein	1,866666667	1,45E-15	1,39	0,028367347
TA04420	hypothetical protein	2,003333333	1,85E-15	1,6	0,032
TA04450	protein disulphide isomerase, putative	1,813333333	3,92E-15	2,08	0,040784314
TA13990	hypothetical protein	1,916666667	4,26E-15	2,12	0,040769231
TA20550	hypothetical protein	1,79	4,56E-15	2,15	0,040566038
TA06670	hypothetical protein	1,78	4,83E-15	2,17	0,040185185
TA08220	hypothetical protein	1,81	7,83E-15	2,49	0,045272727
TA09655	hypothetical protein	1,9	7,94E-15	2,49	0,044464286
TA20910	aminopeptidase n, putative	1,763333333	9,28E-15	2,6	0,045614035
TA02865	Tpr-related protein family member, putative	1,743333333	1,01E-14	2,66	0,045862069
TA08715	bacterial histone-like protein, putative	1,803333333	1,08E-14	2,69	0,04559322
TA16555	hypothetical protein	1,756666667	1,29E-14	2,86	0,047666667

TA11010	hypothetical protein, conserved	1,796666667	1,30E-14	2,87	0,04704918
TA09805	Theileria-specific sub-telomeric protein, SVSP family	1,763333333	1,40E-14	2,97	0,047903226
TA13390	hypothetical protein	1,74	1,55E-14	3,08	0,048888889
TA16685	polymorphic antigen precursor-like protein, putative	1,806666667	1,56E-14	3,08	0,048125
TA08705	elongation factor 1-gamma, putative	1,75	1,70E-14	3,2	0,049230769
TA08580	hypothetical protein	1,713333333	1,73E-14	3,21	0,048636364
TA18145	hypothetical protein	1,8	1,92E-14	3,28	0,048955224

2.4. Full list of genes down-regulated from sporozoite to macroschizont obtained by Rank Product (n=30; FDR≤0.05)

Gene ID	Annotation	FC	RP score	EE	FDR
TA14205	hypothetical protein	-	4,96E-31	0	0
		5,756666667			
TA19275	hypothetical protein	-	1,34E-30	0	0
		5,586666667			
TA03755	sporozoite surface antigen	-4,87	7,63E-27	0	0
TA07435	Sfil-subtelomeric fragment related protein family member, putative	-4,63	1,60E-25	0	0
TA17055	hypothetical protein	-	2,19E-23	0	0
		4,113333333			
TA03805	hypothetical protein	-	1,24E-22	0,01	0,001666667
		4,193333333			
TA05870	roptry-associated protein, putative	-	1,61E-22	0,01	0,001428571
		4,116666667			
TA08520	hypothetical protein	-3,93	6,67E-22	0,02	0,0025
TA21080	Map2 kinase, putative	-3,94	6,71E-22	0,02	0,002222222
TA16420	hypothetical protein	-3,81	7,55E-22	0,02	0,002
TA03475	hypothetical protein	-	2,89E-21	0,03	0,002727273
		3,763333333			
TA18005	hypothetical protein	-	3,21E-21	0,03	0,0025
		3,756666667			
TA14210	hypothetical protein	-	7,35E-21	0,04	0,003076923
		3,596666667			
TA21400	hypothetical protein	-	1,11E-20	0,04	0,002857143
		3,603333333			
TA03290	hypothetical protein	-3,72	1,71E-20	0,04	0,002666667
TA21395	hypothetical protein	-	2,74E-20	0,05	0,003125
		3,573333333			
TA07305	hypothetical protein	-3,57	5,91E-20	0,06	0,003529412
TA08360	hypothetical protein, conserved	-3,48	6,39E-20	0,06	0,003333333
TA04825	hypothetical protein, conserved	-	1,59E-19	0,1	0,005263158
		3,543333333			
TA07630	hypothetical protein, conserved	-	1,90E-19	0,11	0,005
		3,563333333			
TA17595	Sfil-subtelomeric fragment related protein family member, putative	-	2,60E-19	0,11	0,005238095
		3,506666667			
TA16155	hypothetical protein	-	3,58E-19	0,12	0,005454545
		3,313333333			
TA19675	hypothetical protein	-	4,28E-19	0,12	0,005217391
		3,406666667			
TA05340	hypothetical protein, conserved	-3,49	4,62E-19	0,12	0,005

TA11285	hypothetical protein	-3,35	5,02E-19	0,12	0,0048
TA04820	hypothetical protein, conserved	-	5,26E-19	0,12	0,004615385
		3,396666667			
TA11905	hypothetical protein	-	6,01E-19	0,13	0,004814815
		3,353333333			
TA04795	hypothetical protein	-	9,29E-19	0,13	0,004642857
		3,356666667			
TA16660	hypothetical protein, conserved	-3,46	9,66E-19	0,13	0,004482759
TA13045	hypothetical protein, conserved	-	1,09E-18	0,14	0,004666667
		3,423333333			

2.5. Full list of genes up-regulated from merozoite to piroplasm obtained by Rank Product (n=24; FDR≤0.05).

Gene ID	Annotation	FC	RP score	EE	FDR
TA03870	hypothetical protein	4,273333	6,20E-33	0	0
TA09365	hypothetical protein	3,736667	3,21E-29	0	0
TA13885	hypothetical protein	3,206667	3,80E-27	0	0
TA09050	hypothetical protein	3,36	4,69E-27	0	0
TA20315	Theileria-specific hypothetical protein	3,226667	1,46E-26	0	0
TA05015	hypothetical protein	2,896667	2,92E-25	0	0
TA07130	hypothetical protein	2,576667	3,23E-24	0	0
TA05025	hypothetical protein	2,5	6,93E-24	0	0
TA05525	Sfil-subtelomeric fragment related protein family member, putative	2,703333	1,23E-23	0	0
TA03730	cysteine proteinase precursor, tacP	1,986667	2,49E-20	0,03	0,003
TA02715	hypothetical protein	2,043333	2,65E-20	0,03	0,002727
TA13515	hypothetical protein, conserved	1,88	1,75E-19	0,06	0,005
TA15095	Tpr-related protein family member, putative	1,973333	3,84E-19	0,07	0,005385
TA11810	hypothetical protein	1,803333	7,10E-19	0,11	0,007857
TA17605	Sfil-subtelomeric fragment related protein family member, putative	1,896667	1,10E-18	0,15	0,01
TA08875	hypothetical protein	1,85	1,65E-18	0,15	0,009375
TA17500	Sfil-subtelomeric fragment related protein family member, putative	1,763333	8,41E-18	0,29	0,017059
TA09675	hypothetical protein	1,713333	1,43E-17	0,33	0,018333
TA07305	hypothetical protein	1,646667	3,20E-17	0,43	0,022632
TA09205	leucine carboxyl methyltransferase, putative	1,683333	6,56E-17	0,56	0,028
TA03750	cysteine proteinase precursor, tacP	1,61	1,11E-16	0,65	0,030952
TA09450	Sfil-subtelomeric fragment related protein family member, putative	1,723333	1,52E-16	0,7	0,031818
TA08370	hypothetical protein, conserved	1,55	3,59E-16	0,9	0,03913

TA03115	Tash-like protein, putative	1,543333	5,48E-16	1,02	0,0425
----------------	-----------------------------	----------	----------	------	--------

2.6. Full list of genes down-regulated from merozoite to piroplasm obtained by Rank Product (n=20; FDR≤0.05).

Gene ID	Annotation	FC	RP score	EE	FDR
TA20985	hypothetical protein	-2,09333	9,02E-25	0	0
TA14665	hypothetical protein	-2,12667	4,33E-24	0	0
TA14945	hypothetical protein	-1,91	2,39E-23	0	0
TA20590	phosphoenolpyruvate carboxykinase, putative	-1,70667	7,06E-22	0,01	0,0025
TA11915	hypothetical protein	-2,22667	1,10E-20	0,02	0,004
TA14680	hypothetical protein	-1,57667	1,02E-19	0,05	0,008333
TA02750	pepsinogen, putative	-1,61333	1,14E-19	0,05	0,007143
TA12045	membrane protein family member, putative	-1,49667	3,55E-18	0,2	0,025
TA17865	Theileria-specific integral membrane protein, putative	-1,67	8,41E-18	0,29	0,032222
TA05760	rhopty-associated protein, putative	-1,4	1,49E-17	0,34	0,034
TA08220	hypothetical protein	-1,3	1,52E-17	0,34	0,030909
TA03860	pepsinogen, putative	-1,38333	2,83E-17	0,41	0,034167
TA04460	Tpr-related protein family member, putative	-1,35333	3,90E-17	0,45	0,034615
TA08345	hypothetical protein	-1,42333	5,32E-17	0,53	0,037857
TA11905	hypothetical protein	-1,72333	1,15E-16	0,66	0,044
TA05045	hypothetical protein	-1,39667	1,18E-16	0,66	0,04125
TA13535	hypothetical protein	-1,37333	2,57E-16	0,8	0,047059
TA17365	(subtelomeric) ABC-transporter protein family member, putative	-1,37333	3,06E-16	0,83	0,046111
TA20325	integral membrane protein, putative	-1,29333	3,33E-16	0,86	0,045263
TA03840	hypothetical protein	-1,23	4,14E-16	0,93	0,0465

2.7. Full list of genes up-regulated from piroplasm to sporozoite obtained by Rank Product (n=57; FDR≤0.05).

Gene ID	Annotation	FC	RP score	EE	FDR
TA17485		3,556666667	3,00E-27	0	0
TA09505	Sfil-subtelomeric fragment related protein family member, putative	3,316666667	5,31E-25	0	0
TA10650	hypothetical protein	3,266666667	3,06E-24	0	0
TA16700	Sfil-subtelomeric fragment	3,146666667	3,59E-24	0	0

	related protein family member, putative				
TA17346		3,116666667	3,62E-24	0	0
TA11400	Sfil-subtelomeric fragment related protein family member, putative	3,286666667	4,06E-24	0	0
TA11940	hypothetical protein	3,21	1,29E-23	0	0
TA03290	hypothetical protein	2,96	1,30E-22	0,01	0,00125
TA18750	Theileria-specific hypothetical protein	3,13	1,48E-22	0,01	0,001111111
TA11405	subtelomeric sfi-fragment-related protein family member, putative	3,116666667	4,59E-22	0,02	0,002
TA12045	membrane protein family member, putative	2,706666667	4,68E-22	0,02	0,001818182
TA18540	hypothetical protein	2,83	1,48E-21	0,02	0,001666667
TA17550	Theileria-specific subtelomeric protein, SVSP family	2,806666667	9,16E-21	0,02	0,001538462
TA16020	hypothetical protein	2,9	1,08E-20	0,02	0,001428571
TA03755	sporozoite surface antigen	2,63	2,53E-20	0,04	0,002666667
TA14995	hypothetical protein	2,493333333	6,97E-20	0,06	0,00375
TA05540	Theileria-specific subtelomeric protein, SVSP family	2,653333333	9,16E-20	0,06	0,003529412
TA17375	polymorphic antigen precursor, putative	2,67	9,56E-20	0,06	0,003333333
TA12195	subtelomeric sfi-fragment-related protein family member, putative	2,68	6,10E-19	0,1	0,005263158
TA15555	hypothetical protein, conserved	2,626666667	6,15E-19	0,1	0,005
TA09510	Sfil-subtelomeric fragment related protein family member, putative	2,52	8,00E-19	0,11	0,005238095
TA05580	Theileria-specific subtelomeric protein, SVSP family	2,563333333	2,36E-18	0,14	0,006363636
TA17555	Theileria-specific subtelomeric protein, SVSP family	2,44	3,22E-18	0,17	0,007391304
TA18535	hypothetical protein	2,416666667	1,17E-17	0,27	0,01125
TA19275	hypothetical protein	2,283333333	1,67E-17	0,31	0,0124
TA13350	hypothetical protein	2,466666667	2,03E-17	0,33	0,012692308
TA11915	hypothetical protein	2,446666667	2,99E-17	0,38	0,014074074
TA12250	hypothetical protein	2,123333333	4,13E-17	0,47	0,016785714
TA02735	Sfil-subtelomeric fragment related protein family member, putative	2,15	4,47E-17	0,5	0,017241379
TA05555	Theileria-specific subtelomeric protein, SVSP family	2,356666667	5,30E-17	0,52	0,017333333
TA19800	hexokinase 1, putative	1,92	5,38E-17	0,52	0,016774194
TA03150	Tash1-like protein, putative	2,37	1,12E-16	0,63	0,0196875
TA12450	hypothetical protein	1,913333333	1,54E-16	0,68	0,020606061
TA12370	hypothetical protein, conserved	1,883333333	2,36E-16	0,76	0,022352941
TA13595	hypothetical protein	1,993333333	2,84E-16	0,83	0,023714286
TA09435	Theileria-specific subtelomeric protein, SVSP family, putative	2,346666667	3,01E-16	0,84	0,023333333
TA08520	hypothetical protein	2,1	3,14E-16	0,86	0,023243243

TA12945	hypothetical protein, conserved	2,196666667	3,47E-16	0,88	0,023157895
TA12840	hypothetical protein	2,176666667	5,47E-16	1,01	0,025897436
TA03805	hypothetical protein	2,036666667	6,87E-16	1,1	0,0275
TA06060	hypothetical protein, conserved	1,723333333	7,64E-16	1,13	0,027560976
TA12140	subtelomeric sfi-fragment-related protein family member, putative	2,28	8,55E-16	1,15	0,027380952
TA11910	hypothetical protein	2,24	9,38E-16	1,22	0,028372093
TA03615	hypothetical protein	1,923333333	9,70E-16	1,22	0,027727273
TA03130	Tash(AT)-like protein, putative	2,253333333	1,13E-15	1,31	0,029111111
TA10735	GATA-specific transcription factor, putative	2,013333333	1,21E-15	1,32	0,028695652
TA16040	Theileria-specific subtelomeric protein, SVSP family, putative	2,14	1,29E-15	1,33	0,028297872
TA09360	hypothetical protein, conserved	2,096666667	1,32E-15	1,36	0,028333333
TA12760	hypothetical protein	2,02	2,59E-15	1,72	0,035102041
TA13965	hypothetical protein	1,94	5,89E-15	2,23	0,0446
TA12275	Sfil-subtelomeric fragment related protein family member, putative	1,973333333	6,69E-15	2,37	0,046470588
TA17540	Theileria-specific subtelomeric protein, SVSP family	2,1	6,73E-15	2,37	0,045576923
TA04830	hypothetical protein, conserved	1,633333333	6,85E-15	2,37	0,044716981
TA14205	hypothetical protein	2,15	7,00E-15	2,37	0,043888889
TA04895	hypothetical protein, conserved	1,946666667	9,88E-15	2,68	0,048727273
TA04470	hypothetical protein	1,696666667	9,92E-15	2,68	0,047857143
TA05455	ABC transporter, putative	1,996666667	1,16E-14	2,84	0,049824561

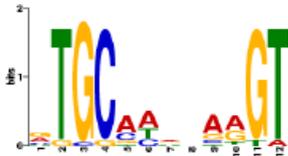
2.8. Full list of genes down-regulated from piroplasm to sporozoite obtained by Rank Product (n=35; FDR≤0.05).

Gene ID	Annotation	FC	RP score	EE	FDR
TA16900	Tpr-related protein family member, putative	-	1,61E-29	0	0
		5,216666667			
TA08535	Tpr-related protein family member, putative	-	1,79E-29	0	0
		5,166666667			
TA08540	Tpr-related protein family member, putative	-	2,01E-28	0	0
		5,066666667			
TA16895	Tpr-related protein family member, putative	-	5,61E-27	0	0
		4,836666667			
TA16870	Tpr-related protein family member, putative	-	2,73E-26	0	0
		4,663333333			
TA13885	hypothetical protein	-	1,72E-25	0	0
		4,453333333			
TA09365	hypothetical protein	-3,72	2,79E-24	0	0
TA16765	Tpr-related protein family member, putative	-4,1	6,27E-24	0	0
TA16875	Tpr-related protein family member, putative	-3,99	1,45E-23	0,01	0,001111111
TA06635	Tpr-related protein family member, putative	-	1,56E-23	0,01	0,001
		3,993333333			
TA03870	hypothetical protein	-3,59	2,98E-23	0,01	0,000909091
TA16790	Tpr-related protein family member, putative	-3,82	4,90E-23	0,01	0,000833333

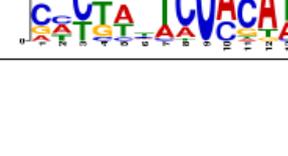
TA06745	Tpr-related protein family member, putative	- 3,403333333	4,74E-22	0,02	0,001538462
TA16795	Tpr-related protein family member, putative	- 3,333333333	6,72E-22	0,02	0,001428571
TA05605	hypothetical protein, conserved	-3,13	1,78E-21	0,02	0,001333333
TA06740	Tpr-related protein family member, putative	-3,24	2,13E-21	0,02	0,00125
TA20315	Theileria-specific hypothetical protein	- 3,186666667	4,04E-21	0,02	0,001176471
TA06895	Tpr-related protein family member, putative	-3,16	4,89E-21	0,02	0,001111111
TA06535	hypothetical protein	- 3,173333333	1,18E-20	0,02	0,001052632
TA09675	hypothetical protein	- 2,883333333	3,31E-20	0,05	0,0025
TA06900	Tpr-related protein family member, putative	- 2,913333333	3,68E-20	0,06	0,002857143
TA16685	polymorphic antigen precursor-like protein, putative	- 2,943333333	4,37E-20	0,06	0,002727273
TA18875	Sfil-subtelomeric fragment related protein family member, putative	- 2,773333333	8,49E-20	0,06	0,002608696
TA06580	Tpr-related protein family member, putative	- 2,806666667	1,01E-19	0,07	0,002916667
TA06880	Tpr-related protein family member, putative	-2,71	1,74E-19	0,07	0,0028
TA08285	Tpr-related protein family member, putative	-2,77	2,41E-19	0,08	0,003076923
TA05015	hypothetical protein	-2,64	3,38E-19	0,09	0,003333333
TA14645	Tpr-related protein family member, putative	- 2,653333333	4,44E-19	0,09	0,003214286
TA21045	hypothetical protein	- 2,626666667	4,49E-19	0,09	0,003103448
TA06330	Tpr-related protein family member, putative	- 2,566666667	1,11E-18	0,11	0,003666667
TA08370	hypothetical protein, conserved	- 2,503333333	1,17E-18	0,11	0,003548387
TA02865	Tpr-related protein family member, putative	- 2,476666667	1,70E-18	0,13	0,0040625
TA13515	hypothetical protein, conserved	-2,55	2,11E-18	0,13	0,003939394
TA09050	hypothetical protein	- 2,453333333	2,88E-18	0,17	0,005
TA08360	hypothetical protein, conserved	- 2,463333333	2,97E-18	0,17	0,004857143

Appendix 3 - *T. annulata* DNA motifs identified in Chapter 3

3.1. Top 3 motifs identified by MEME in the 5' intergenic regions of the 100 most up-regulated genes during macroschizont to merozoite stage differentiation in *T. annulata*.

	Motif	Width	E-value	Sites
1		14	1.3e-009	50
2		14	9.6e-016	40
3		20	5.5e-009	10

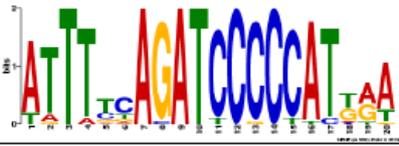
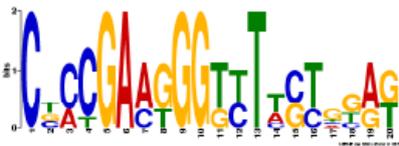
3.2. Top 3 motifs identified by MEME in 5' intergenic regions of the 100 most down-regulated genes during differentiation of the macroschizont to merozoite stage in *T. annulata*.

	Motif	Width	E-value	Sites
1		20	1.0e-155	55
2		20	2.4e-001	17
3		20	1.5e+001	17

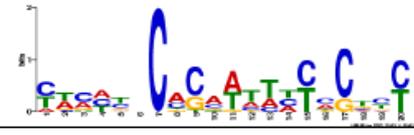
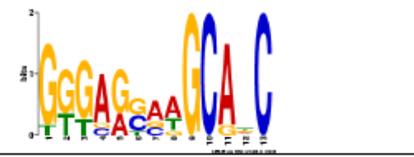
3.3. Top 3 motifs identified by MEME in the 3' intergenic regions of the 100 most up-regulated genes during macroschizont to merozoite stage differentiation in *T. annulata*.

	Motif	Width	E-value	Sites
1		20	6.7e-005	17
2		20	1.6e-001	11
3		20	1.1e+001	5

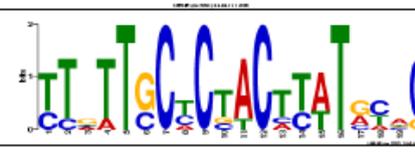
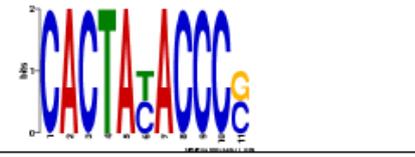
3.4. Top 3 motifs identified by MEME in 3' intergenic regions of the 100 most down-regulated genes during differentiation of the macroschizont to merozoite stage in *T. annulata*.

	Motif	Width	E-value	Sites
1		20	6.6e-118	30
2		10	1.2e-006	24
3		20	2.5e-005	7

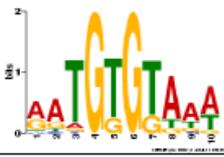
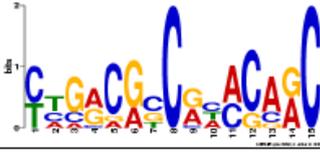
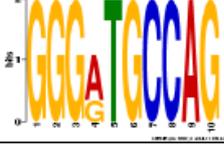
3.5. Top 3 motifs identified by MEME in 5' intergenic regions of ApiAP2 genes of *T. annulata* in macroschizont to merozoite stage.

	Motif	Width	E-value	Sites
1		20	6.1e-002	23
2		13	1.1e+000	9
3		10	2.7e+003	20

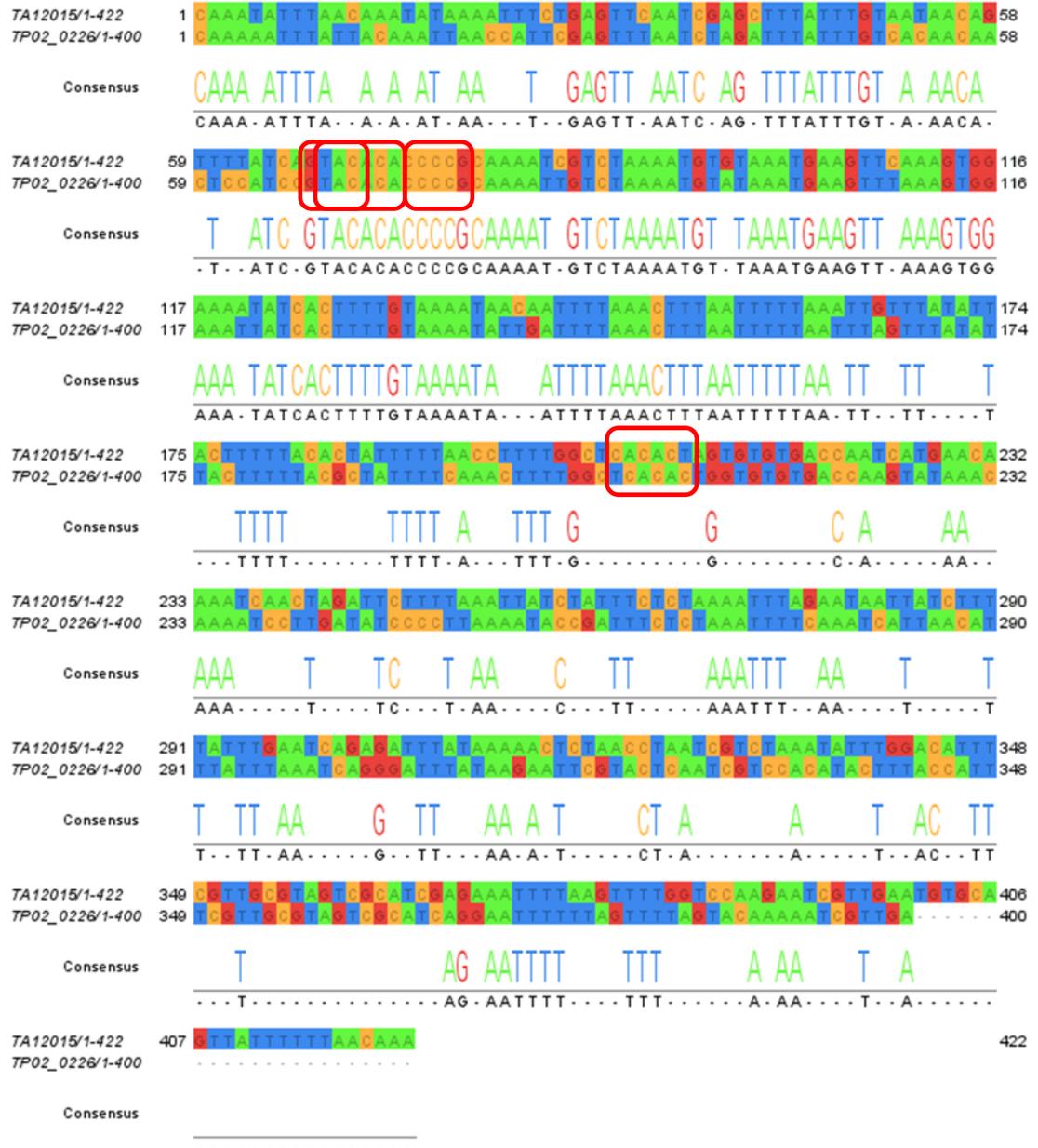
3.6. Top 3 motifs identified by MEME in 5' intergenic regions of other transcription factors of *T. annulata* in macroschizont to merozoite stage.

	Motif	Width	E-value	Sites
1		10	7.6e-003	14
2		20	1.6e+003	6
3		11	2.6e+005	2

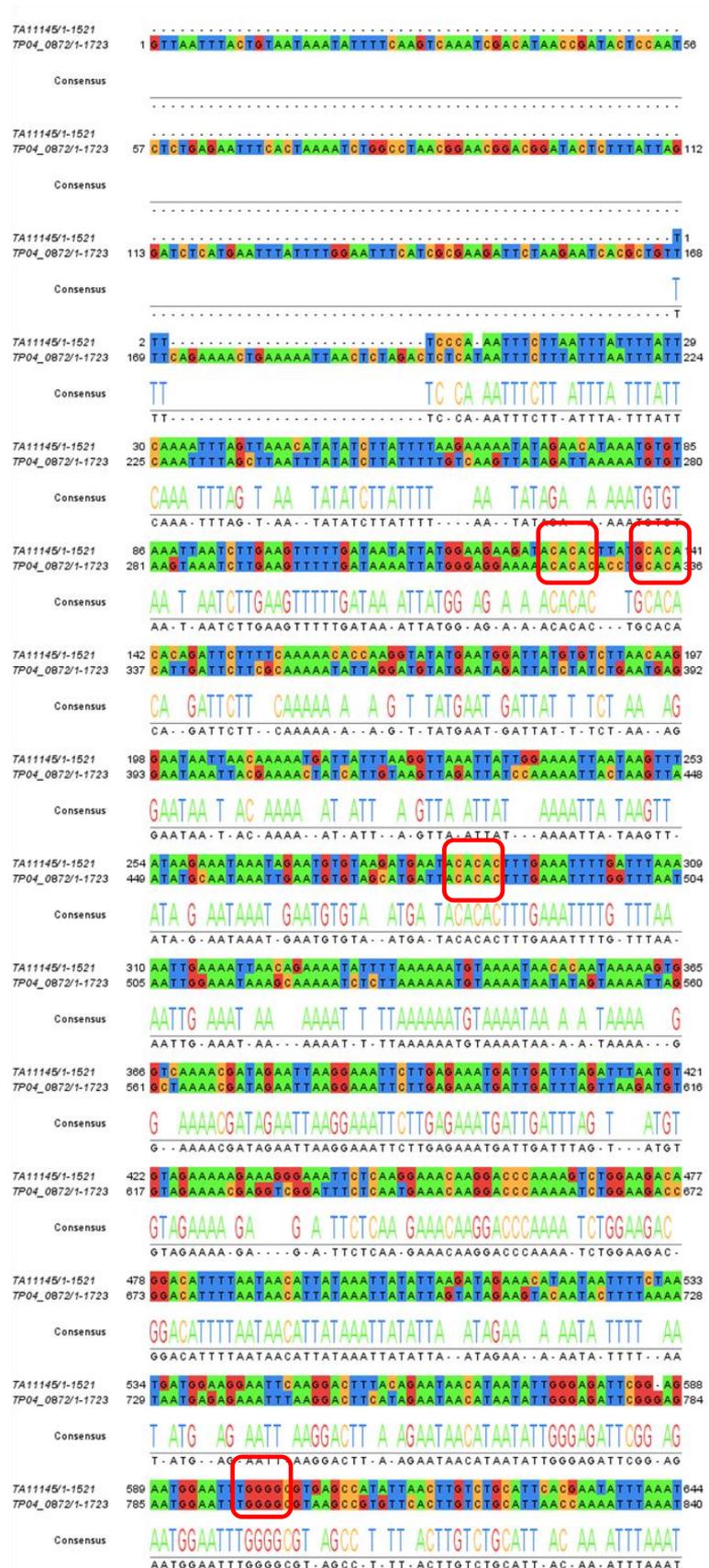
3.7. Top 3 motifs identified by MEME in 5' intergenic regions of *T. annulata* constitutive genes from macroschizont to merozoite stage

	Motif	Width	E-value	Sites
1		10	2.3e-015	98
2		15	3.0e+00 3	10
3		10	4.2e+00 5	3

3.8. Full alignment of upstream sequences of *T. annulata* and *T. parva* orthologues of TA12015. Highlighted in red ApiAP2 motifs. Highlighted in red binding motifs predicted for up-regulated AP2 factors (TA11145 - (A)CACAC(A); TA12015 -G/C-box; TA13515 - GTAC; TA16485 - TCTA(C/T)A).



3.9. Full alignment of upstream sequences of *T. annulata* and *T. parva* orthologues of TA11145. Highlighted in red binding motifs predicted for up-regulated AP2 factors (TA11145 - (A)CACAC(A); TA12015 -G/C-box; TA13515 - GTAC; TA16485 - TCTA(C/T)A).



TA11145/1-1521 645 AACTGCTTAACTGATTCTCTCAAGTGTAAACATTAGTTTATACACTATATTTTC 700
 TP04_0872/1-1723 841 ACGCCCTTTGGCACTTACTCTCTCAAGTGTAAACATTAGTTTATAAACTATATTTTC 896
 Consensus A TT CACT A TCTCTCAAGTGTAAACATTAGTTTATA ACTATATTTTC
 A - - - - TT - - CACT - A - TCTCTCAAGTGTAAACATTAGTTTATA - ACTATATTTTC

TA11145/1-1521 701 TATAATTAATATTTAAATTTTCACTCTAGTATCCATTTTATTAGGTGTCGTGAAAC 756
 TP04_0872/1-1723 897 TATAATTTTATATTTAAATTTTCACTCTACTATCCCTTTTATTAGGTGTCGTGAAAC 952
 Consensus TATA TT ATATTTAAATTTT ACCTCTA TATCC TTTTATTAGGTGTCGTGAAAC
 TATA - TT - ATATTTAAATTTT - ACCTCTA - TATCC - TTTTATTAGGTGTCGTGAAAC

TA11145/1-1521 757 ACCTCTGTTCCCTTCACTTGACACTTTTAAATTTCTCTACACATGTTAAACCTTATA 812
 TP04_0872/1-1723 953 ACCTCGGCTCCCTTACTCGACACTTTTAAATTTCTCTACACATGTTAAACCTTACA 1008
 Consensus ACCTC G TCCTT ACT GACACTTTTAA TTCTCGTACACATC TTAACCTTA A
 ACCTC - G - TCCTT - ACT - GACACTTTTAA - TTCTCGTACACATC - TTAACCTTA - A

TA11145/1-1521 813 CAATAAAAAGGATATAAACAAATTTGATAATTAATTTAAACCACAAAGATACGGTAT 868
 TP04_0872/1-1723 1009 CAATAAAAAGGATATAAACAACTCGTTAATTAATTTAAACCACAAAGATATGATAT 1064
 Consensus CAATAAAAAGGATATAAACAA T G TAATTAATTTAAACCACAA GATA G TAT
 CAATAAAAAGGATATAAACAA - T - G - TAATTAATTTAAACCACAA - GATA - G - TAT

TA11145/1-1521 869 TTGTAAGCCCTTTTATTTCTGTGTGACCAAGATGGCGACTTCCCTTTTATAGATT 924
 TP04_0872/1-1723 1065 TTGTACGACCTTTTATTTCTGTGTGATCATGATGGCGACATACAAATTTACAGATT 1120
 Consensus TTGTA G CCTTTTATT CTGTGTGA CA GATGGCGAC T C TTT AGATT
 TTGTA - G - CCTTTTATT - CTGTGTGA - CA - GATGGCGAC - T - C - - - TTT - AGATT -

TA11145/1-1521 925 TCCAAAAACATTTCTTCAATATTTTAAATTTAAATATTTTGGCAAAGTATCACGTC 980
 TP04_0872/1-1723 1121 TCTAAAAACCTCTCTTCAATAATTTTAAATTTAAATATTTTGGAAACATTCACGTC 1176
 Consensus TC AAAAC T TCTTCAT ATTTTAAATTTAAATATTTTGG AA TCAC G
 TC - - AAAAC - T - TCTTCAT - ATTTTAAATTTAAATATTTTGG - - AA - - TCAC - - G

TA11145/1-1521 981 TAATCCC AAAATATTTATAATTTAATTTTGGTATTTCTCTTAAATGCCAATCAC 1036
 TP04_0872/1-1723 1177 CAATTC CCAAAAATATTTAGACTATTAATTTTGTAGTCTCTTAAATGCCAATCAC 1232
 Consensus AATTC CCAAAAATATTTA A TATTAATTTTGG TA TTCTTTAATGC A AC
 - AATTC CCAAAAATATTTA - A - TATTAATTTTGG - TA - TTCTTTAATGC - A - AC

TA11145/1-1521 1037 TGTAAATATTTTGTATTTAATTTTAAACCTGAGGCGCACTTGTCAAATTTGG 1092
 TP04_0872/1-1723 1233 TGTAAATATTTTGTATTTAATTTTAAACCTGAGGCGCACTTGTCAAATTTGG 1288
 Consensus TGTAAATATTTTGTATTTAAT T TAAACCTG AGC GA ATTGTTCAAAA TGG
 TGTAAATATTTTGTATTTAAT - - T - TAAACCTG - AGC - GA - ATTGTTCAAAA - TGG

TA11145/1-1521 1093 ATCATAACACATTCCTTTGAAATGATAGTTGCTAATATTTTAAAGATGACCTTT 1148
 TP04_0872/1-1723 1289 ATCATAACACATTCCTTTGAAATGATAGTTGCTAATATTTTAAAGATGACCTTT 1344
 Consensus ATCATA ACATTC CTTTGAA GATAGTT C A TATTTTAAAGATGA C TTT
 ATCATA - ACATTC - CTTTGAA - GATAGTT - C - A - TATTTTAAAGATGA - C - TTT

TA11145/1-1521 1147 TAAAACTCATTAACCCCGAGATTGTTACAGATAACACACTTTAAATCTTGT 1202
 TP04_0872/1-1723 1345 TAACATTTTCAACCTCCTGGATTGCTTACAGATAACACACTTTAAATCTTGT 1399
 Consensus TAA A T A A C C C G GATTGG TTACAGATAACACACT TAAATCTTGT
 TAA - A - T - - - A - A - C - C - C - G - GATTGG - TTACAGATAACACACT - TAAATCTTGT

TA11145/1-1521 1203 ACAAAGTTAAGCTGGAAGGGTAAATGTCGAGTACGTAACCCCTCATTCTTAAACAA 1258
 TP04_0872/1-1723 1400 ACAAAGTTAAGCTGGAAGGGTCAATGACCGTAAAGTACCCCTCATTCTTAAACAA 1455
 Consensus ACAAAGTTA CTGGAAGGGT AATG CG GTA GTACGCCCTCA TT T AAAAC A
 ACAAAGTTA - CTGGAAGGGT - AATG - CG - GTA - GTACGCCCTCA - TT - T - AAAAC - A

TA11145/1-1521 1259 TACA - - ATAATTAATGAACCTTTAATAATAACACAAAACAGTGTCAATTGATGAC 1311
 TP04_0872/1-1723 1456 CTACACTATAATTAATGACCTTTTATCATAAAACACCAAAATGTCAATCTTTGAC 1511
 Consensus TACA ATAATTAATGA CTTT AT ATAA ACA A CAA TGTC AAT TGAC
 - TACA - ATAATTAATGA CTTT - AT - ATAA - ACA - A - CAA - TGTC AAT - - - TGAC

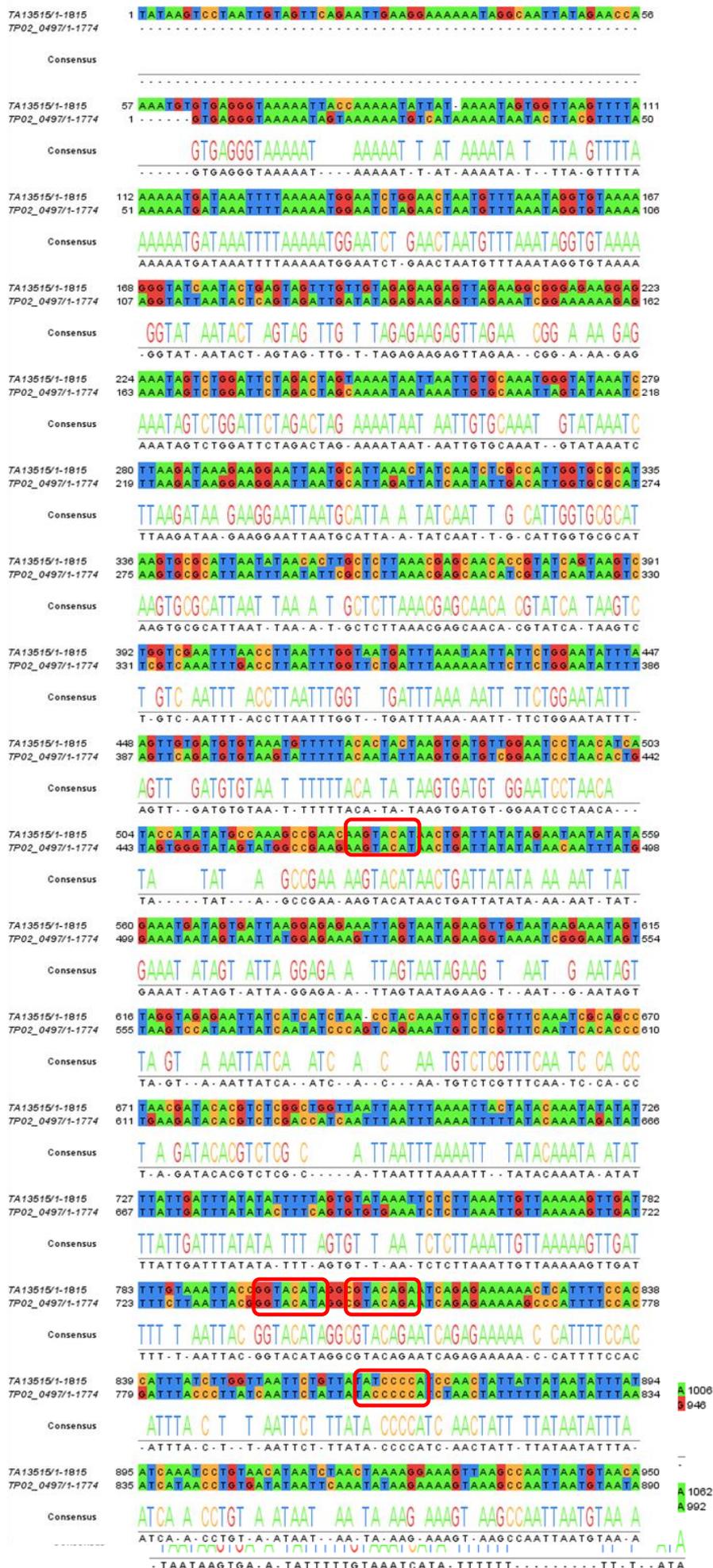
TA11145/1-1521 1312 AAATGTACACATTAACACACACTTATGAGTATCAACACACAATGTATTAATAG 1367
 TP04_0872/1-1723 1512 AAATGTACACATTAACACACTTATGAGTATCAACACACAATGTATTAATAG 1567
 Consensus AAATGATACACATA GA ACAC G TATGAGTATCAACACACA ATGTATTAATAG
 AAATGATACACATA - GA - ACAC - G - TATGAGTATCAACACACA - ATGTATTAATAG

TA11145/1-1521 1368 ACAAACCTTACTGAGAAAATAACTATATAGAAATAAACTAACCTACAATTCCTGAG 1423
 TP04_0872/1-1723 1568 ACAAACCTTACTGAGAAAATTACTATATGGAATAAACTTATTACCACAAATAAAGTG 1623
 Consensus ACAAACCTTACT GAAA T ACTATAT GAATAAA T A C A A G G
 ACAAACCTTACT - - GAAA - T - ACTATAT - GAATAAA - - T - A - - - C - A - - - A - - G - G

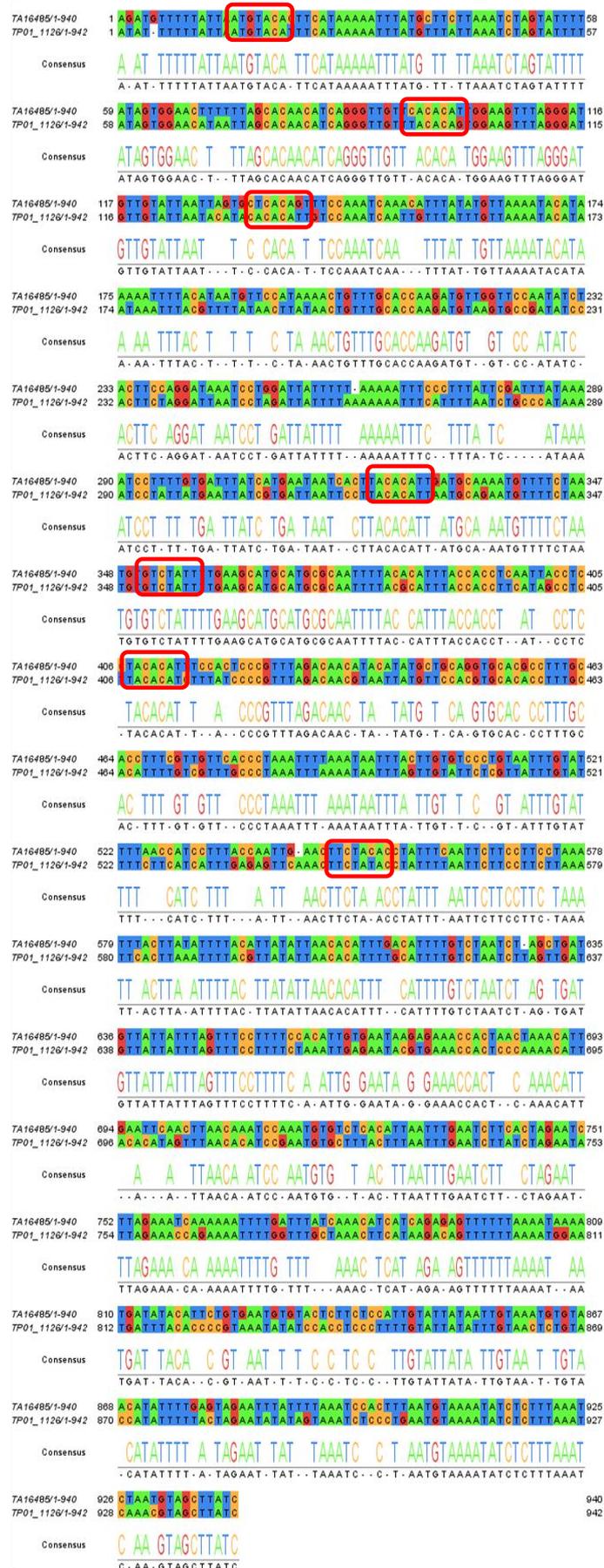
TA11145/1-1521 1424 AAATTTAATTTAAATA - - TAAAACCTGAGTTAATTTTATCAACATAGAAATGAA 1477
 TP04_0872/1-1723 1624 AAATTTAATTTAATTTGCTAAAACCTGAGTTAATTTTATCAACATAGAAATGAA 1679
 Consensus AAATTTAATTTAA T AAAACCTGAA TTTAATTT ATCAA A GAAATGAA
 AAATTTAATTTAA - T - - - AAAACCTGAA - TTTAATTT - ATCAA - A - - GAAATGAA

TA11145/1-1521 1478 ATTTGAGGTTTTGTAATTTAAATCTCAAGTTTATATATTTATC 1521
 TP04_0872/1-1723 1680 ATTTAAGGTTTTAGGATCTTAAATTTCAATTTTAAATATTTATA 1723
 Consensus ATTT AGGTTTT AT TAAAT TCAA TTTA ATATTTAT
 ATTT - AGGTTTT - - - AT - TAAAT - TCAA - TTTA - ATATTTAT -

3.10. Full alignment of upstream sequences of *T. annulata* and *T. parva* orthologues of TA13515. Highlighted in red AP2 motifs.



3.11. Full alignment of upstream sequences of *T. annulata* and *T. parva* orthologues of TA16485. Highlighted in red binding motifs predicted for up-regulated AP2 factors (TA11145 - (A)CACAC(A); TA12015 -G/C-box; TA13515 - GTAC; TA16485 - TCTA(C/T)A).



Appendix 4 – qRT-PCR results of (A)CACAC(A) binding TaApiAP2s from Chapter 4

- 4.1. Comparison of relative expression levels of TA11145 vs TA07100 in D7 and D7B12 cell lines. Expression of TA07100 is significantly higher than TA11145 in both cell lines, and expression of TA11145 relative to TA07100 at Day 0 is significantly lower in the D7B12 cell line (9.5 log₂ fold reduction) compared to D7 cell line (5 log₂ reduction) - 16 fold change absolute.

