GENOMIC AND POPULATION GENETIC STUDIES ON *THEILERIA ANNULATA*

WILLIAM WEIR, BVMS, MRCVS

For the degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY
of
GLASGOW

Division of Veterinary Infection and Immunity
University of Glasgow Veterinary School

September 2006

Author's declaration

This thesis is entirely the product of my own efforts. The work on which it is based was my own, except where specifically stated in the text and in the acknowledgements section. It has not been previously submitted to any university for the award of a degree. The following publications include work contained in this thesis -

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

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

William Weir

September 2006

Abstract

Tropical theileriosis, caused by the tick-transmitted protozoan *Theileria annulata*, is a major disease of cattle in many regions of the developing world. Current research is directed towards developing a sub-unit vaccine, and it is therefore important that genetic diversity in field populations of the parasite is investigated and quantified. The recently completed genome sequence provided an opportunity to develop a panel of genetic markers for population studies and also enabled the identification of novel antigen genes.

The genome was bioinformatically screened to identify micro- and mini-satellite loci, several of which were PCR amplified from a series of diverse parasite stocks in order to characterise their polymorphism and to determine their species-specificity. A panel of ten markers were selected for population genetic studies and were used to genotype laboratory-maintained cell lines and clonal stocks of *T. annulata* isolated from different countries. Cell lines comprised a multiplicity of genotypes, while clonal stocks showed evidence of a single haploid genome. Preliminary population genetic analysis revealed a large amount of genotypic diversity both between and within countries and indicated that the parasite population is geographically sub-structured. Comparison of a limited number of stocks isolated in different countries demonstrated that genetic differentiation between populations positively correlates with intervening physical distance. A low standard index of association (I^S_A) suggested that the population in Tunisia is in linkage equilibrium, indicating that the parasite possesses a panmictic (randomly mating) population structure. To confirm these findings, a large number of field isolates from Tunisia and Turkey were analysed (n = 305). This supported the earlier finding that geographical sub-structuring separates panmictic populations and an almost identical amount of genetic differentiation between countries was evident ($F_{ST} = 0.05$). Limited linkage disequilibrium was observed in some populations and this was attributed to several factors including inbreeding and the Wahlund effect, caused by putatively immigrant sub-populations. A similar multiplicity of infection was demonstrated in vaccinated and unvaccinated animals and the immunising genotype did not appear to establish in the field population. Multiplicity of infection was instead shown to positively correlate with the host age in several sampling locations.

The genome of *T. annulata* was compared with that of *T. parva* to identify gene families under the influence of positive selection using mean family inter-genomic non-synonymous to synonymous substitution rates ($d_N d_S$). Codon usage between the species and between several life-cycle stages within *T. annulata* was shown to be virtually invariant and independent of the $d_N d_S$ distribution. In addition to a subset of merozoite genes, which were predicted to be antigens on the basis of their motif signature, a sub-telomeric gene family (SVSP) and a family of parasite-encoded host nuclear genes (TashATs) showed evidence of positive selection between the species. An allele-sequencing approach was taken to verify these predictions which indicated that, in general, the TashAT genes are under the effect of purifying selection while two SVSP genes were shown to be highly variable, however there was no firm evidence of positive selection. One of the merozoite antigen candidates showed evidence of both positive immune selection and balancing selection. Consequently, further studies are indicated to assess whether this gene has value as a vaccine candidate.

Table of Contents

Chapter Six

List of Tables

List of Figures

Abbreviations and symbols

A	adenine **or** alanine
A+	untemplated adenine
ANCOVA	analysis of co-variance
Aromo	aromaticity score
Batan2	'Batan deu'; Tunisian cell line
B-cells	B-lymphocytes
BLAST	basic local alignment search tool
bp	base pairs
C	cytosine
C9	*T. annulata* clone used for genome sequencing
C-terminus	carboxyl terminus of protein
$CD4^+$	cluster of differentiation 4 - positive
$CD8^+$	cluster of differentiation 8 - positive
cDNA	complimentary DNA
CDS	coding sequence
CI	confidence interval
cM	centimorgan
COA	correspondence analysis
CSP	*P. falciparum* circumsporozoite protein
CTL	cytotoxic T-lymphocytes
D	Nei's genetic distance
DNA	deoxyribonucleic acid
d_N	rate of non-synonymous substitutions
$d_N d_S$	non-synonymous to synonymous substitution rate
d_S	rate of synonymous substitutions
ECF	East Coast Fever
ELISA	enzyme linked immunosorbent assay
EN_c	effective number of codons
EST	expressed sequence tag
FAM	blue fluorochrome used to label PCR primer
F_{ST}	standardised measure of genetic differentiation among populations
G	guanine
GBS	group B Streptococcus
GC_{3s}	frequency of guanine or cytosine nucleotides at the third position in synonymous codons
GC_{skew}	skew in the frequency of guanine and cytosine nucleotides
GeneDB	www.genedb.org; *T. annulata* online genomic resource
GPI	glucose phosphate isomerase
GPI-anchor	glycosyl-phosphatidylinositol anchor
Gravy	hydrophobicity score
GS500	Genescan™ 500 size markers
GSS	genome sequence survey
G_{ST}'	estimator of F_{ST} which is independent of number of samples
H_e	estimated heterozygosity
HKA test	Hudson-Kreitman-Aguade test
H_S	within sample gene diversity

IFAT	immunofluorescent antibody test
INF-γ	interferon-gamma
I_S^A	standardised index of association
ISCOM	immune stimulating complex
Jed4	'Jedeida quatre'; Tunisian cell line
k	nucleotide differences
kb	kilobase (one thousand bases)
kDa	kilodalton
L	95 % confidence limit (linkage analysis)
L_{aa}	amino acid length
L_{PARA}	95 % confidence limit (linkage analysis / parametric test)
L_{MC}	95 % confidence limit (linkage analysis / Monte Carlo simulation)
LB	Luria Broth
LD	linkage disequilibrium
LE	linkage equilibrium
lsa-1	*P. falciparum* liver stage antigen 1
mAb	monoclonal antibody
Mb	megabase (one million bases)
mero1	merozoite candidate gene 1 (TA13810)
mero2	merozoite candidate gene 1 (TA20615)
MHC	major histocompatibility complex
mg kg^{-1}	milligrams per kilogram
MK test	McDonald-Kreitman test
ML	maximum likelihood
MLG	multilocus genotype
MPSA	major merozoite / piroplasm surface antigen
mRNA	messenger RNA
MS	micro- and mini-satellite
MSA-2	*P. falciparum* merozoite surface antigen 2
n	number of samples
NF-κB	nuclear factor-kappa B
NJ tree	neighbour joining tree
NK cells	natural killer cells
NLS	nuclear localisation signal
NO	nitric oxide
N-terminus	amino terminus of protein
ORF	open reading frame
p	*p* value; statistical significance
p67	*T. parva* 67 kDa sporozoite protein
p104	*T. parva* 104 kDa microneme-rhoptry antigen
p105	105 kDa pre-cursor of NF-κB
p150	*T. parva* 150 kDa microsphere antigen
PBM	peripheral blood mononuclear cells
PBS	phosphate buffered saline
PCA	principal component analysis
PCR	polymerase chain reaction
PCR-RFLP	polymerase chain reaction – restriction fragment length polymorphism
PCV	packed cell volume

PEST	peptide motif comprising proline (P), glutamic acid (E), serine (S) and threonine (T) amino acid residues
pfs48/45	*P. falciparum* gamete surface protein
Pfu	*Pyrococcus furiosus* polymerase
PIM	*T. parva* polymorphic immunodominant molecule
PQ	amino acids proline (P) and glutamine (Q)
r^2	goodness of fit
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
ROX	red fluorochrome used to label Genescan™ size standards
rRNA	ribosomal RNA
RSCU	relative synonymous codon usage
RT-PCR	reverse transcription-polymerase chain reaction
S	serine **or** number of segregating sites
SCF	standard chromatogram format
SE	standard error
SLAC	single likelihood ancestor counting
SLAG-1	*T. lestoquardi* sporozoite antigen 1
SNP	single nucleotide polymorphism
SPAG-1	*T. annulata* sporozoite antigen 1 (TA03755)
$SuAT_1$	*T. annulata* encoded host-nuclear protein (TA03135)
SVSP1	SVSP sequencing candidate gene 1 (TA16025)
SVSP2	SVSP sequencing candidate gene 2 (TA17485)
SVSP	sub-telomeric variable secreted protein
T	thymine **or** threonine
TaA_2	Ankara A_2 stock of *T. annulata*
TaMS1	*T. annulata* merozoite antigen 1 (TA17050)
Taq	*Thermus aquaticus* polymerase
Tar	*T. annulata* repeat; multi-copy locus found in *T. annulata*
TashHN	*T. annulata* schizont host-nuclear protein (TA20090)
TaSP	*T. annulata* surface protein (TA17315)
*Ta*Tu	*T. annulata* / Tunisia; reference system used by Ben Miled
T-cells	T-lymphocytes
TMD	transmembrane domain
TNF-α	tumour necrosis factor alpha
TpMS1	*T. parva* merozoite surface antigen 1
Tpr	*T. parva* repeat; multi-copy locus found in *T. parva*
URL	uniform resource locator; website address
UTR	untranslated region
V_D	variance of mismatch values (linkage analysis)
VSG	variant surface glycoprotein
VNTR	variable number of tandem repeats
ºC	degrees Celsius
µl	microlitres
µm	micrometres
π	nucleotide diversity
θ	genetic variation (θ_S / θ_π) **or** estimator of F_{ST}

Acknowledgements

I would like to thank my supervisors, Professors Andy Tait and Brian Shiels for providing excellent guidance and support throughout this project. I appreciate the encouragement and constructive criticism of Professor Shiels while I am particularly grateful to Professor Tait for his constant enthusiasm and for providing me with the opportunity to make the transition from veterinary practice to laboratory research.

I thank Dr Jane Kinnaird, Dr David Swan and Ms Sue McKellar for providing invaluable assistance and instruction in the ways of molecular biology during a steep learning curve. I am indebted to Dr Arnab Pain from the Sanger Institute in Cambridge for making *T. annulata* genomic resources available to me and also to Dr Frank Katzer from the Centre for Tropical Veterinary Medicine in Edinburgh for supplying parasite material.

I would like to thank Professor Mohamed Darghouth for providing parasite material and for hosting me at his laboratory in the École Nationale de Médecine Vétérinaire de Sidi Thabet, Tunisia. I am especially grateful to Dr Mohamed Gharbi who, in addition to assisting with parasite DNA preparation and PCR reactions, provided hospitality and encouraged me to consume local delicacies, considered inedible in most regions of the world.

I am grateful to Dr Tülin Karagenç and Professor Hasan Eren at Adnan Menderes University, Aydın for their generosity during my visit to Turkey. I am particularly indebted to Dr Karagenç for the provision of parasite material together with an extensive amount of data and for assistance in DNA purification. I am also grateful for the company of Murat Hosgor and Huseyin Bilgic who introduced me to the invigorating experience of the Turkish bath.

I thank my assessors Dr Barbara Mable and Dr Annette Macleod for useful advice and suggestions. Over the last few years I have enjoyed the company of my colleagues, Tiggy Grillo, Erica Packard and Libby Redman who have graciously endured my patter.

I would like to acknowledge the financial support of Glasgow University Veterinary School.

Finally, I would like to express my gratitude to my partner, Hayley, who like the unpainted windows and unmown lawn, I have neglected in recent times. The unreserved support of Hayley together with my parents, has allowed me to undertake this project and it is to these three people whom I dedicate this thesis.

CHAPTER ONE

INTRODUCTION

1.1. Introduction

Tropical or Mediterranean theileriosis is an economically important bovine disease, widespread in North Africa, Southern Europe, India, the Middle East and Asia, placing an estimated 200 million cattle at risk (Purnell 1978). The disease is caused by the protozoan parasite *Theileria annulata*, affects domestic cattle (*Bos taurus* and *Bos indicus*) and Asian buffalo (*Bubalus bubalis*) and is transmitted by several species of ixodid ticks of the genus *Hyalomma*. The severity of the condition varies between limited clinical reactions in endemically stable areas to high morbidity in endemically unstable regions where mortality among calves can be high. Pathogenesis is primarily due to proliferation of infected leucocytes and anaemia.

Emerging factors, including the development of drug resistance in *T. annulata* (personal communication, M. Darghouth) and the potential for alteration in vector habitat through climate change (Estrada-Pena 2003), may have a profound impact on the epidemiology of *T. annulata*. However, before such epidemiological fluctuations can be studied in detail, the basic underlying population structure of the parasite must be elucidated and appropriate genotyping tools developed. Traditionally, control of the disease is aimed at limiting contact between the cattle and the tick, and includes regular use of acaricides. However, with the use of acaricides being curtailed by vector resistance (Musoke *et al.* 1996) the long-term role of these chemicals for prophylaxis is in doubt. In addition, there are logistical and quality control issues with current attenuated cell line vaccines. Consequently, there is a need for improved vaccines, which can overcome these difficulties. It is the combination of disease control problems and the potential for change in disease epidemiology that stimulated the research presented in this thesis.

1.2. *Theileria* species of veterinary significance

The genus *Theileria* encompasses a number of species of tick-transmitted parasitic protozoa that occur in domestic livestock and other mammals. They are classified in the phylum apicomplexa along with *Babesia*, *Eimeria*, *Plasmodium* and *Toxoplasma* (Levine 1985). There are five species of *Theileria* that cause clinical disease in cattle, the two most economically important clinical presentations being tropical theileriosis, caused by

T. annulata, and East Coast Fever, a highly fatal disease of cattle in East and Central Africa, caused by *T. parva*. It has been suggested that the buffalo was the original host for an ancestor of both *Theileria* spp., although it is unclear whether divergence occurred preceding or following cattle adaptation (Uilenberg 1981). In both cases, the disease is usually mild in buffalo, but pathogenic to cattle, particularly non-indigenous breeds.

The worldwide distribution of both *T. annulata* and *T. parva* is shown graphically in Figure 1.1., along with that of *T. sergenti*, a pathogenic species of *Theileria* in the Far East. All the *Theileria* species of veterinary significance together with their associated diseases are summarised in Table 1.1. and are discussed below.

1.2.1. *Theileria annulata*

T. annulata was first described in 1904 in Transcaucasian cattle (Dschunkowsky and Luhs 1904) and was named *Piroplasma annulatum*. The parasite was reclassified as *Theileria annulata* following identification of the schizont stage in the life-cycle (Bettencourt *et al.* 1907), although a multitude of confusing terminology was in use throughout endemic regions of the world in the first two decades of the 20th century. Following recognition of tropical theileriosis as a major constraint to agricultural development in North Africa, Sergent and co-workers at the Institut Pasteur d'Algerie undertook a concerted study on the disease in Algeria, producing much of the primary basic research on the parasite over the following 30 years.

In North Africa, tropical theileriosis occurs from Morocco in the west through to Egypt in the east. The range extends into to sub-Saharan Africa in Sudan and Eritrea, where it may overlap with the range of *T. parva* (see Figure 1.1.). Cattle from many countries in Southern Europe including Spain, Portugal, Italy and Greece are endemically infected in their Southern regions. The disease extends through Turkey, the Near and Middle East, Central Asia and India, to Southern Russia and Northern China in the Far East. The extent of this disease reflects the availability of suitable habitat for the ixodid tick vector (Purnell 1978). In the Far East, the yak (*Bos grunniens*) is also known to be highly susceptible to infection. In contrast to the other main bovine piroplasm *Babesia* spp., which are transmitted trans-ovarially from one tick generation to the next, *Theileria* spp. are transmitted trans-stadially by two- and three-host ticks. As many as fifteen *Hyalomma* species of tick have been identified as vectors for tropical theileriosis (Robinson 1982). *T. annulata* is maintained in nature by a cattle – tick – cattle cycle, although infection may

Table 1.1. Clinical diseases of domestic ruminants caused by *Theileria* spp.

The genus *Theileria* encompasses several species of tick-transmitted parasitic protozoa that cause disease in domestic livestock. Of the six species of *Theileria* that cause clinical disease in ruminants, the two most economically important presentations occur in cattle and are commonly known as tropical theileriosis, caused by *T. annulata*, and East Coast Fever (ECF), caused by *T. parva*.

Table 1.1. Clinical diseases of domestic ruminants caused by *Theileria* spp.

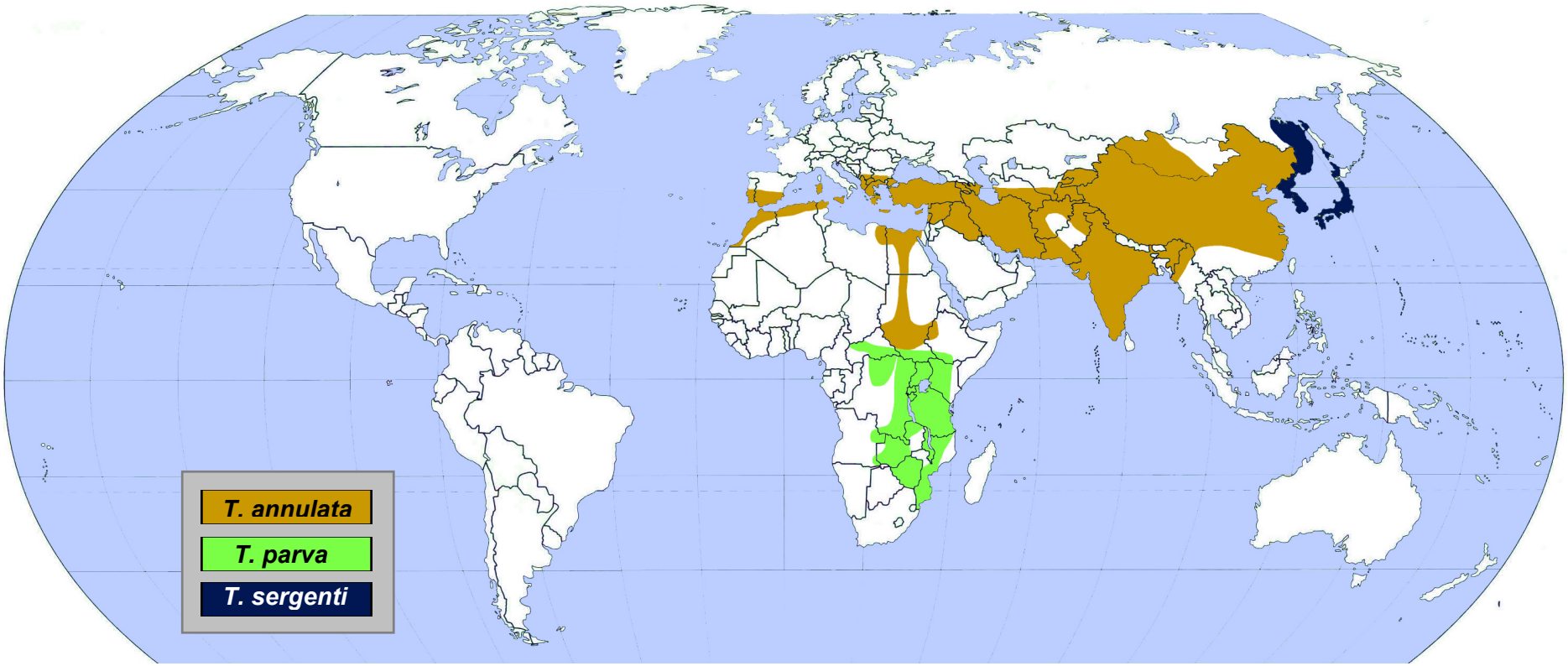
Disease	Clinically affected host	Parasite species	Transformation of host cells	Vector	Distribution
Tropical theileriosis	cattle and yak (*Bos grunniens*)	*Theileria annulata*	√	*Hyalomma* spp.	North Africa, Southern Europe, Near and Middle East, Central Asia, India and Northern China
East Coast Fever, Corridor disease and Zimbabwe theileriosis	cattle	*Theileria parva*	√	*Rhipicephalus appendiculatus* and other *Rhipicephalus* spp.	Eastern and Southern Africa
T. taurotragi infection	cattle (first identified in eland)	*Theileria taurotragi*	√	*Rhipicephalus* spp.	Eastern and Southern Africa
Turning sickness	cattle	*Theileria parva* or *Theileria taurotragi*	-	*Rhipicephalus* spp.	Eastern and Southern Africa
T. mutans infection	cattle	*Theileria mutans*	-	*Amblyomma* spp.	Eastern Africa
Malignant ovine / caprine theileriosis	sheep and goats	*Theileria lestoquardi*	√	*Hyalomma* spp.	Mediterranean basin, Sudan, Western and Central Asia and India
T. sergenti [*] infection	cattle, asiatic buffalo (*Bubalus bubalis*)	*Theileria sergenti*[*]	-	*Haemaphysalis* spp.	Eastern Asia, including Japan

[*] *T. sergenti* is classified within the group *T. buffeli / orientalis*, the taxonomy of which is unresolved

Figure 1.1. Distribution of major *Theileria* spp. of cattle

The worldwide distribution of both *T. annulata* and *T. parva* along with that of *T. sergenti*, a pathogenic species of *Theileria* in the Far East is presented opposite. *T. annulata* extends from Southern Europe and North Africa into sub-Saharan Africa in Sudan and Eritrea, where it may overlap with the range of *T. parva*. Tropical theileriosis extends through Turkey across to Southern Russia and Northern China in the Far East, reflecting the availability of suitable habitat for the tick vector.

Figure 1.1. Distribution of major *Theileria* spp. of cattle



also be transmitted when an infected male accidentally detaches while feeding and reattaches to another animal (Pipano and Shkap 2006).

In India alone, tropical theileriosis is a major health problem for ten million exotic and crossbred cattle with an expected economic loss of US$ 800 million per annum (Wilkie *et al.* 1998). The financial impact of disease has recently been documented in Tunisia, examining (a) direct costs due to veterinary intervention, (b) the cost of mortality and (c) reduced productivity (Tisdell *et al.* 1999; Gharbi *et al.* 2006). More than half of the cost was accounted for by sub-clinical infection. A cost-benefit analysis of using a cell line vaccine demonstrated that use of the vaccine would be economical even if the price were to dramatically increase from € 5 to € 73 per dose.

1.2.2. *Theileria parva*

T. parva is principally transmitted by the brown ear tick, *Rhipicephalus appendiculatus*, and in addition to causing East Coast Fever (ECF), it is also responsible for three other bovine syndromes in sub-Saharan Africa – i.e. Corridor disease, Zimbabwe theileriosis and Turning sickness.

ECF has been endemic in Eastern Africa for a long time with the disease spreading widely in the area during the early part of the 20th century following European colonisation with imported, susceptible cattle (Norval *et al.* 1992). When uncontrolled, the disease may cause over 90 % mortality in such stock with clinical signs of pyrexia, lymphadenopathy, pulmonary oedema and death, that are similar to those described for tropical theileriosis (Lawrence *et al.* 2006b). A small proportion of animals may survive, but recovery is prolonged and incomplete.

Corridor disease clinically resembles acute ECF and is a result of ticks transmitting buffalo-derived strains of *T. parva* (Lawrence *et al.* 2006a). However, this form of the parasite is not well adapted to cattle where failure to develop to the piroplasm stage makes cattle – tick – cattle transmission unlikely. Turning sickness is an aberrant form of *T. parva* infection, where an accumulation of infected lymphoid cells in the cerebral vasculature precipitates an afebrile nervous condition (Lawrence and Williamson 2006b).

1.2.3. *Theileria sergenti*

A further *Theileria* species responsible for significant disease in cattle is *T. sergenti*, first described by Yakimoff and Dekhtereff in 1930 in the Vladivostok area of Eastern Siberia as *Gonderia sergenti*. Although the name *T. sergenti* has also been applied to a parasite of

sheep by Wenyon in 1926 (Morel and Uilenberg 1981), it is in common usage to describe the causative agent of severe clinical infections of cattle in Japan, Eastern Russia and Eastern China. The parasite's current taxonomic status is as a member of the *T. buffeli / orientalis* group, however, the name *T. sergenti* will be used for simplicity throughout this thesis.

In an attempt to clarify the phylogenetic status of parasites known as *T. buffeli* / *orientalis* group, a molecular study using small subunit RNA gene sequences (Chansiri *et al.* 1999) identified two groups of *Theileria* spp., which were strongly supported by bootstrap analysis. The pathogenic *T. annulata*, *T. parva* and *T. taurotragi* clustered within one group along with newly isolated *Theileria* parasites of cervidae, whereas *T. sergenti* was found within the other group, which consisted of non-pathogenic species of *T. buffeli*. All members of this predominantly non-pathogenic group, including *T. sergenti* are benign, in that they are unable to transform and direct proliferation of host cells.

1.2.4. *Theileria lestoquardi*

T. lestoquardi is a highly pathogenic ovine and caprine parasite, and is generally accepted as being the only species of economic significance in these hosts (Lawrence 2006). It is transmitted by *Hyalomma anatolicum anatolicum* and occurs in South-eastern Europe, Northern Africa, Southern Russia and the Middle East. The parasite has been shown to be antigenically closely related to *T. annulata* (Leemans *et al.* 1997), although *T. lestoquardi* is incapable of infecting cattle (Leemans *et al.* 1999). Experiments in sheep indicate that *T. lestoquardi* infection protects against subsequent *T. annulata* infection (Leemans *et al.* 1999) and although prior infection with *T. annulata* does not prevent infection from *T. lestoquardi* sporozoites, it does protect against the major clinical effects. The pathogenesis of malignant ovine theileriosis is similar to that of the pathogenic *Theileria* spp. in cattle.

1.2.5. Other pathogenic *Theileria* spp.

Theileria mutans is widespread in sub-Saharan Africa and on some Caribbean islands (Lawrence and Williamson 2006a) with a life-cycle typical of *Theileria* but with parasite proliferation mainly occurring during the piroplasm stage of the life-cycle, rather than the schizont stage. In Eastern Africa the parasite may cause severe clinical infection and mortality, resulting from profound anaemia caused by intravascular haemolysis in combination with a heavy piroplasm parasitaemia. *Theileria taurotragi* is transmitted by several *Rhipicephalus* species and often occurs as a mixed infection with *T. parva* and

along with that species, it is considered to be one of the causes of Turning sickness (de Vos *et al.* 1981).

1.3. Life-cycle of *Theileria annulata*

1.3.1. Cattle stages

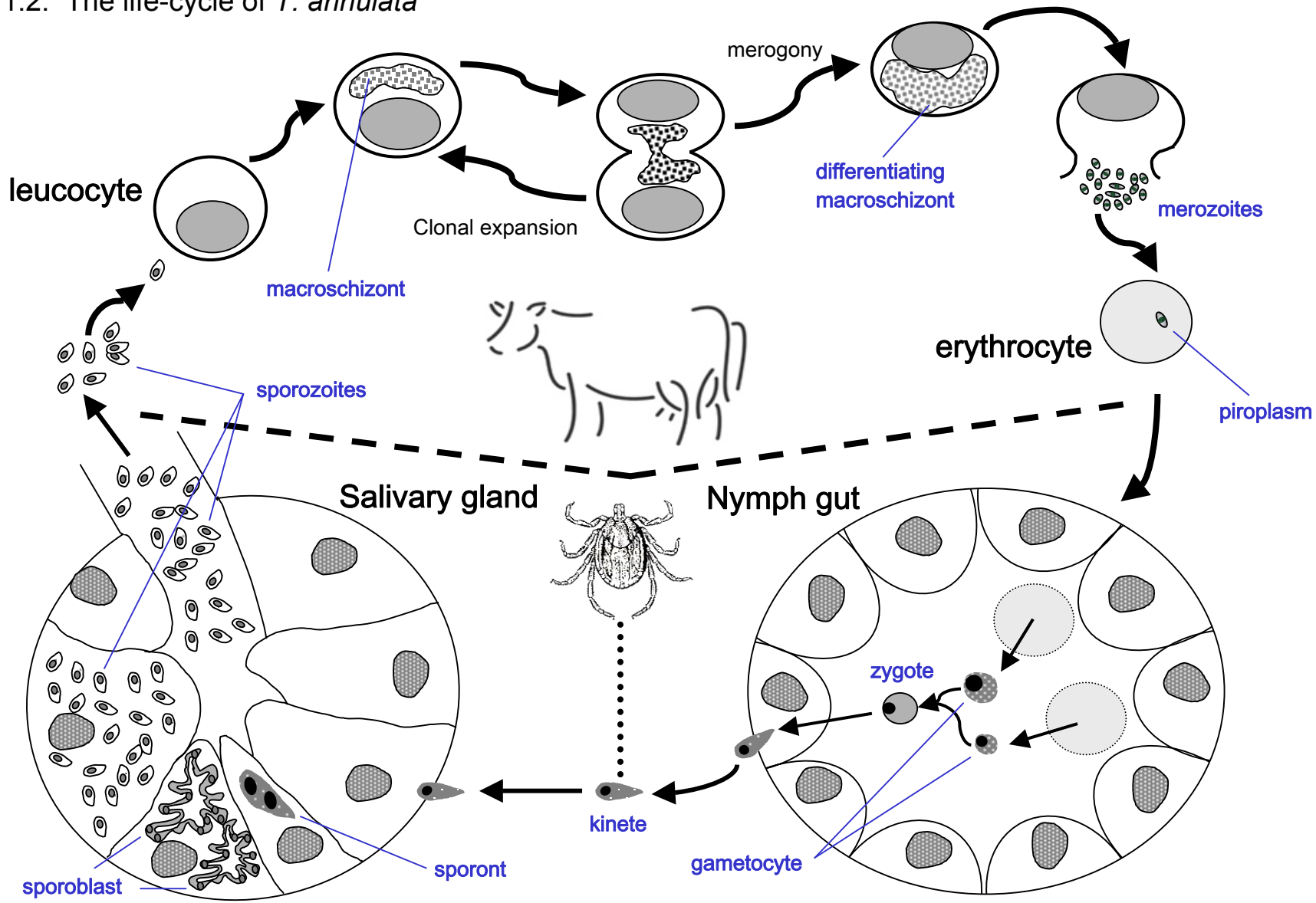
The tick vector infects the cattle host while taking a blood meal. Occasionally, infection may be contracted congenitally by calves; consequently all cases occurring during the first week of life may be considered congenital (Levine 1985). The infective stage, the sporozoite, is an oval shaped body measuring 1μm in length. Sporozoites are formed in the salivary gland of the adult tick and following inoculation, while the tick is taking a blood meal, invade leucocytes and develop into the uninucleate trophozoite (Figure 1.2.). Different bovine leucocyte populations are known to be preferentially infected by *T. annulata* and *T. parva*. This was demonstrated by *in vitro* experiments that compared infectivity of sporozoites from each parasite species against several purified bovine cell types - monocytes, T-cells and major histocompatibility complex (MHC) class II positive and negative sub-populations of cells (Glass *et al.* 1989). *T. annulata* was found to preferentially infect macrophage-type cells and MHC class II positive cells. T-cells showed a low level of infection, while the level of infection in MHC class II negative cells was negligible. In marked contrast, *T. parva* preferentially infected T-cells but was unable to infected monocyte-type cells. A similar study demonstrated B-cells are infected much more efficiently by *T. annulata* than *T. parva* (Spooner *et al.* 1989).

For *Theileria* spp. parasite proliferation and stage differentiation takes place in the host cell cytoplasm, as opposed to a parasitophorous vacuole as occurs with some other apicomplexans such as *Plasmodium* and *Toxoplasma* (Mehlhorn and Schein 1984). The trophozoite develops within the host cell to form the macroschizont, a multinucleate stage of the parasite that induces host cell division and synchronises its own proliferation to that of the infected leucocyte by binding to the mitotic spindle. In culture, macroschizonts contain, on average, twelve nuclei, each nucleus measuring about 1.5 μm in diameter (Mehlhorn and Schein 1984). The proliferating infected leucocytes quickly produce a large population of infected cells. Proliferation initially occurs in the lymph nodes but subsequently infected cells may be found in the bloodstream or invade a range of tissues and organs giving the disease a ‘cancer-like’ quality, i.e. uncontrolled host cell reproduction and ‘metastasis’. Macroschizont infected cells may be detected in smears from superficial lymph nodes and the liver, 7 to 28 days post-infection.

Figure 1.2. The life-cycle of *T. annulata*

The cattle host is infected by a feeding tick, which inoculates sporozoites that invade myeloid leucocytes to form the multinucleate macroschizont stage. Infected cells proliferate with a proportion of macroschizonts differentiating into merozoites. Following breakdown of the host cell membrane, merozoites invade erythrocytes to become piroplasms, the infective stage for the feeding tick nymph. Following a blood meal, gametocytes develop in the tick gut, which fuse to form zygotes and then transform into motile kinetes; in an otherwise haploid state, the zygote and kinete are the only stages where the parasite is considered diploid. Kinetes migrate to the salivary gland of the vector where after several rounds of asexual multiplication, sporozoites are formed. The related life-cycle of the tick vector is presented in Figure 1.5.

Figure 1.2. The life-cycle of *T. annulata*



In vivo, a proportion of macroschizonts undergoes a process known as merogony where the schizont becomes enlarged followed by the production of a large number of uninucleate merozoites, which lie free in the host cell cytoplasm. An electron-microscopic study in *T. parva* demonstrated the process by which the syncytial differentiating macroschizont packages the prescribed assortment of organelles into free merozoites (Shaw and Tilney 1992). In general, eukaryotic cells are unable to synthesise organelles *de novo*, so it is critical that the full complement of organelles is present in each merozoite. To ensure this occurs, the organelles are bound to the nuclear envelope, which then associates with the schizont plasma membrane. Following this, merozoites detach from the main syncytial body in an orderly manner known as 'budding' (Shaw and Tilney 1992). An analogous process is believed to occur in *T. annulata*. The merozoite contains a single nucleus, one or two mitochondria, a microneme and between three and six rhoptries (Mehlhorn and Schein 1984; Levine 1985).

Following the breakdown of the host cell plasma membrane, mature merozoites are liberated into the bloodstream, which then invade red cells and form intra-erythrocytic merozoites, known as piroplasms. These may be detected one to three days after the first appearance of macroschizonts (8 – 10 days following infection), and may persist for years in infected cattle (Pipano and Shkap 2006). Piroplasms are generally spherical, oval or comma shaped bodies bounded by a single-layered cell plasma membrane (Mehlhorn and Schein 1984). In culture, it was demonstrated that piroplasms undergo a division process characteristic of merogony, i.e. karyokinesis is followed by cytokinesis (Conrad *et al.* 1985). Four daughter piroplasms may be produced, with the same features as merozoites generated within leucocytes, however this intra-erythrocytic division is considered to be limited in *T. parva* and *T. annulata*. In these species, red cells are continually invaded by merozoites produced from infected leucocytes residing in immunologically privileged sites (Ilhan *et al.* 1998). It is likely that within the erythrocyte, the process of differentiation to gametes is initiated, pre-adapting the parasite for the tick phase of the life-cycle. It is the parasitised red cell, which is infective to the feeding tick vector.

1.3.2. Vector stages

Between one and four days post-repletion, piroplasms within ingested erythrocytes develop into slender, spindle-shaped microgamonts within the tick gut (Schein *et al.* 1975). Up to four nuclei are formed within these microgamonts along with several flagellum-like appendages, which break away to create filiform microgametes. By this time, spherical stages, termed macrogametes, have also appeared. It is presumed that the micro- and

macro-gametes unite to form a zygote, although this fusion has not been observed (Schein *et al.* 1975; Levine 1985). Zygotes have a vacuole-like centre and appear from five days post-repletion in the gut epithelium. After a further growth period of around 12 – 15 days, motile elongated forms, termed kinetes, are liberated into the haemolymph and migrate to the salivary gland of the nymph (Schein and Friedhoff 1978). The zygote and the kinete are presumed to be the only diploid forms, in the primarily haploid life-cycle of this parasite. Within type II and III acini, kinetes invade tick cells and transform into fission bodies. These structures grow in size as their nucleus divides until the onset of vector moulting when development ceases (Schein and Friedhoff 1978).

The act of feeding activates sporogony and allows parasite development to progress. A study of *Hyalomma anatolicum anatolicum* demonstrated that unfed ticks were incapable of producing infection in experimental calves (Singh *et al.* 1979). Sporozoites appeared in the salivary gland after adults had been feeding for two to three days and were at a maximal level at 72 hours (Singh *et al.* 1979). It has been shown that a temperature of 37 °C and a relative humidity of 95 % is in itself sufficient to stimulate the production of infective sporozoites in infected adult *Hyalomma excavatum* ticks without the need for a blood meal (Samish 1977), suggesting some ticks may be infective before feeding commences. Once sporogony is activated, a series of fissile events results in the production of sporozoites, which are liberated into the tick saliva as the host acinar cells degenerate. It has been estimated that 40,000 sporozoites may be formed from each infected acinus (Young *et al.* 1992).

The molecular processes that govern and control the formation of different stages throughout the life-cycle are not fully understood. However, it has been proposed that the processes of sporogony, merogony (in the leucocyte) and piroplasm differentiation (in the erythrocyte) occur by a morphologically similar mechanism which may be under the control of a single cassette of genes (Shaw and Tilney 1992). A stochastic model of differential gene expression during merozoite formation has been postulated (Shiels 1999).

1.4. The genomes of *T. annulata* and *T. parva*

T. annulata, like *T. parva* has a haploid genome with a brief diploid phase in the tick, when zygotes are formed. The nuclear genome of *T. annulata* is estimated to be 8.35 Mb and is arranged in four chromosomes ranging in length from 1.9 to 2.6 Mb within which 3,792 coding regions have been predicted (Pain *et al.* 2005). Although the genome of *T. parva* is slightly smaller, 8.31 Mb, it is predicted to code for an extra 238 genes. The Sanger

Institute fully sequenced the genome of *T. annulata* in a collaborative project with the University of Glasgow and the Moredun Research Institute while at the same time the genome of *T. parva* was sequenced by the Institute for Genomic Research (TIGR) in Maryland (Gardner *et al.* 2005). A whole-genome shotgun of the *T. annulata* Ankara clone C9 was carried out to '8x' coverage, with shotguns of individual chromosomes performed to a further '2x' coverage. Bacterial artificial chromosome (BAC) end sequencing to '10x' was also employed to further assist assembly (http://www.sanger.ac.uk/Projects/T_annulata/). Orthologues in *T. parva* have been identified for more than 85 % of *T. annulata* genes with the genomes showing strong conservation of synteny outside sub-telomeric regions (Pain *et al.* 2005). Additionally, the presence of a small extra-chromosomal mitochondrial DNA element (6.5 kb in length) has been documented in *T. annulata* and *T. parva* (Hall *et al.* 1990) and the apicoplast genome sequenced for *T. parva* (Gardner *et al.* 2005).

1.5. Clinical signs and pathogenesis

The dose of sporozoites inoculated (Samantaray *et al.* 1980; Preston *et al.* 1992) determines the onset and severity of clinical signs, which are generally around 9 to 25 days after feeding of the infected tick. The initial clinical signs are pyrexia (41 ºC) and lymphadenopathy by which time schizonts may be detected in the liver and the spleen. This is accompanied by an increase in pulse and respiratory rates after which anorexia, ruminal stasis and general paresis ensue. In the peracute form of the disease, susceptible animals die within three to five days following an overwhelming infection. In the acute presentation, the disease lasts for one to two weeks and is very often fatal in adult cattle. This contrasts with the situation in immune cattle, which are resistant to *T. annulata* infection through either passive immunity or following vaccination. In such animals clinical signs are mild or absent and recovery is spontaneous (Pipano and Shkap 2006). In young animals diarrhoea may be observed, although this is likely to be due to intercurrent infection with intestinal pathogens as the immune response is compromised by lymphocytopaenia (Sharpe and Langley 1983).

In a recent study in Saudi Arabia (Omer *et al.* 2003) 62 clinical cases of infection were documented in adult and juvenile Friesian cattle. The most prominent gross pathological features were petechial and ecchymotic haemorrhages involving many mucosal and serosal surfaces in addition to the body fat. Lesions, similar to those occurring in *T. parva*, were observed in animals suffering lethal infection.

Clinical signs are principally considered to be the result of leucoproliferation, primarily associated with destruction and disorganisation of the lymphoid system and anaemia resulting from the rupture of piroplasm-infected erythrocytes. In addition, there is emerging evidence that cytokine production by schizont-infected cells as well as their proliferation is important in disease pathogenesis (Preston *et al.* 1993). It has been shown experimentally that schizont-infected cells disseminate rapidly through lymphoid tissue initially from the lymph node draining the area of inoculation to distant lymph nodes and the spleen and thymus (Forsyth *et al.* 1999). By the late stage of the disease the number of erythrocytes harbouring piroplasms greatly outnumbers macroschizont infected white cells. A prompt and severe leucopaenia accompanies infection (Preston *et al.* 1992) due to a decline in circulating neutrophils and lymphocytes. The Forsyth study also suggested that parasitised mononuclear cells phagocytose piroplasm-infected erythrocytes and thus may play a role in the development of anaemia and tissue damage. Icterus, which is often a feature of disease, is a consequence of destruction of erythrocytes and hepatic damage. Haemolysis is, however, not solely the result of piroplasm infection as cattle experimentally infected with attenuated strains of *T. annulata*, unable to progress to the piroplasm stage, may still suffer anaemia (Darghouth *et al.* 1996a).

It has been proposed that tumour necrosis factor alpha (TNF-α) plays a significant role in the pathogenesis of tropical theileriosis (Preston *et al.* 1993; Brown *et al.* 1995; Forsyth *et al.* 1999). Excessive production of TNF-α by infected bovine macrophages (Preston *et al.* 1993) may be related to leucopaenia and the petechiae and eccyhmoses characteristic of the disease (Forsyth *et al.* 1999). High levels of interferon-γ (INF-γ) found in lethal *T. annulata* infections may induce this increase in TNF-α *in vivo* (Campbell *et al.* 1997a; Campbell *et al.* 1997b).

1.6. Immunity

The ability of indigenous cattle to resist tropical theileriosis, coupled with the fact that cell line vaccination is successful in protecting otherwise susceptible stock, demonstrates that the bovine immune system is capable of mounting an effective response to both initial and subsequent infection. These results suggest that resistant breeds of cattle may possess a degree of innate immunity, while vaccinated exotic stock rely on an acquired response following either vaccination or primary challenge. However, evidence to date indicates that the innate and adaptive bovine immune responses act in concert against both primary and secondary challenge. This contrasts with the situation in *T. parva*, where immunity is afforded principally by the adaptive response of cytotoxic T-cells.

1.6.1. The innate response

In vitro, sporozoites have been shown to preferentially invade and transform monocytic cells and B-cells to a much larger extent than T-cells (Sager *et al.* 1997; Sager *et al.* 1998). Macroschizont-infected mononuclear cells are known to exhibit a messenger RNA (mRNA) cytokine profile similar to uninfected activated macrophages (Preston *et al.* 1999). Additionally, infected cells are able to activate uninfected 'cytostatic' macrophages to produce cytokines (including TNF-α), nitric oxide (NO) and factors that can suppress proliferation of infected cells (Preston and Brown 1988). Natural killer (NK) cells are also induced to destroy macroschizont-infected cells (Preston *et al.* 1983), perhaps as a result of IFN-α secretion. The innate immune response mediated by these two cell types is depicted in Figure 1.3. NK cells may also be induced by cytokines from infected cells to produce IFN-γ, which may reinforce NO production by uninfected macrophages (Preston *et al.* 1999). NO has been shown to inhibit the proliferation of macroschizont-infected cells *in vitro* and induce apoptosis, thus precluding differentiation to the merozoite stage (Richardson *et al.* 1998). In addition, NO has been shown to inhibit sporozoite invasion *in vitro* (Visser *et al.* 1995) and it has been speculated that it may inhibit sporozoite and merozoite invasion *in vivo* (Preston *et al.* 1999).

It has been proposed that direct stimulation of the innate immune response may help to control a low infective dose, preventing the parasite from establishing intra-cellularly and therefore resulting in a sub-clinical infection (Preston *et al.* 1999). However, cattle infected with a larger dose of sporozoites would require the assistance of the adaptive immune response to defend against and eliminate subsequent parasite stages. The interplay between innate and adaptive immunity is also depicted in Figure 1.3.

1.6.2. The adaptive responses

1.6.2.1. Cell mediated immunity

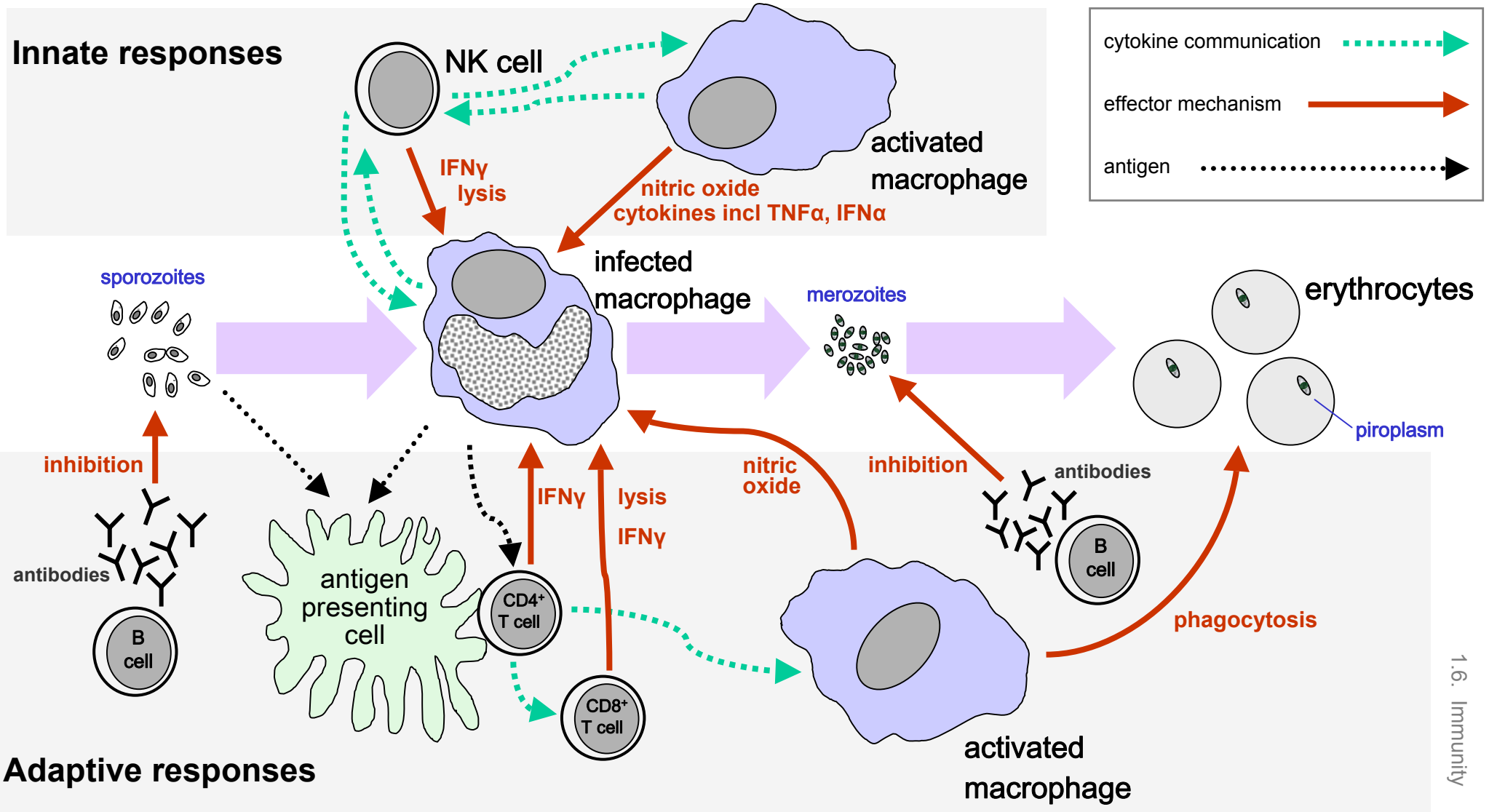
The adaptive response to tropical theileriosis includes stimulation of B-cells and T-cells, although the antigens responsible for inducing a protective response have not yet been elucidated. The mechanism by which infected leucocytes activate macrophages to induce an adaptive response is also unknown. Both $CD4^+$ and $CD8^+$ cell types secrete cytokines including IFN-γ, which assist feeding the autocrine loop between macrophages and NK cells that is involved in innate immunity (Preston *et al.* 1999). This was established when IFN-γ was found to enhance synthesis of TNF-α and NO by macrophages *in vitro* (Preston *et al.* 1993; Visser *et al.* 1995). The NO produced by these activated macrophages acts on

Figure 1.3. The innate and adaptive responses to *T. annulata* infection

The parasite progresses from the sporozoite to the piroplasm stage of infection and may be subject to the action of the immune system at every stage. Natural killer (NK) cells and activated macrophages provide the innate response to primary infection and act upon trophozoite- and macroschizont-infected leucocytes. NK cells lyse infected cells and also modulate activated macrophages by secreting cytokines such as tumour necrosis factor-α (TNF-α), interferon-α (INF-α) and nitric oxide. The adaptive response includes cytotoxic $CD8^+$ T-cells lysing macroschizont infected cells and $CD4^+$ T-cells activating macrophages following stimulation by antigen. Following secondary challenge, the invasive stages (i.e. the sporozoite and merozoite) may be inhibited by antibodies produced by B-cells. Interferon-γ (INF-γ) produced by both $CD4^+$ and $CD8^+$ T-cells acts directly on trophozoite-infected cells; additionally $CD8^+$ T-cells may lyse macroschizont-infected cells, which have survived the earlier responses. Activated macrophages phagocytose piroplasm-infected erythrocytes.

(adapted from Preston *et al.* 1999)

Figure 1.3. The innate and adaptive responses to *T. annulata* infection



infected leucocytes and may additionally act on free merozoites and infected erythrocytes. Macrophages also have the ability destroy infected erythrocytes directly by phagocytosis.

MHC class I molecules, present on the surface of most cells, are thought to signal the infected status of a cell through the presentation of short polypeptides derived from endogenous parasite antigens. $CD8^+$ cytotoxic T-cells are presumed to be important mediators of immunity against *T. annulata* (Ahmed *et al.* 1989) by recognising MHC class I associated antigen. An earlier study demonstrated that during primary infection, cytotoxic cells could only be detected in the circulation and lymph nodes of calves that recovered from infection (Preston *et al.* 1983). A study by Innes *et al.* in 1989, provided evidence that the cytotoxic response to *T. annulata* infection is MHC restricted (Innes *et al.* 1989a). Two groups of cattle were immunised with either an autologous or an allogeneic infected cell line. Cattle inoculated with the allogeneic line developed mild clinical signs while the group inoculated with the autologous cell line exhibited a severe response. In the latter group, cytotoxicity was directed against the parasite, however in the former group, immunised with a foreign bovine cell type, cytotoxicity was first directed against foreign MHC antigens, before a secondary cytotoxic response was mounted against the parasite. All cattle were found to be immune to heterologous sporozoite challenge, confirming the cross-reactivity of the response. While known to lyse macroschizont-infected cells (Innes *et al.* 1989b), $CD8^+$ cytotoxic T-cells are only detectable transiently in acute infection. In contrast, macrophage activity against such infected cells is sustained over a prolonged period. It has been suggested that macrophages may play a more important role in generating the protective immunity elicited by cell line vaccination due to their prolonged activity against infected leucocytes (Preston and Brown 1988; Preston *et al.* 1999). The response to secondary challenge has also been related to the ability of $CD4^+$ memory cells to be rapidly recalled, a cell type that is capable of activating macrophages. Immunity directed against the infected leucocyte is also discussed in Section 1.10.2. in the context of macroschizont antigens.

1.6.2.2. Humoral immunity

Although cattle produce antibodies to parasite antigens during primary infection, immune bovine serum fails to recognise either the surface of the infected leucocyte (Shiels *et al.* 1989) or infected erythrocyte (Hall 1988; Ahmed *et al.* 1988) and for this reason antibodies to these stages are unlikely to play a role in protective immunity. In contrast, immune serum has been shown to block invasion of sporozoites *in vitro* (Gray and Brown 1981; Preston and Brown 1985). *In vivo*, however, sporozoites are apparently only affected by

raised antibody levels following multiple challenge with sporozoites (Preston and Brown 1985). It may be speculated that because sporozoites are only present in the bloodstream transiently, they are only exposed to the immune system for a brief period of time. It is interesting to note that anti-sporozoite antibodies are also able to retard the transformation of trophozoite-infected cells into the proliferating macroschizont stage (Preston and Brown 1985).

Antibodies to merozoites have been demonstrated *in vivo* following primary infection (Irvin and Morrison 1987). Furthermore, it has been shown that immune serum is capable of recognising free merozoites (Ahmed *et al.* 1988) and that complement alone may result in their lysis. However, the latter study also demonstrated there was no inhibitory effect of immune serum on proliferation of infected cells *in vitro*. An experiment performed on *T. parva* demonstrated that the humoral response is unable to combat clinical infection in any detectable way (Muhammed *et al.* 1975). Thus, when naïve animals were transfused with (a) immune serum or (b) globulin prepared from immune serum and then challenged, all cattle developed fatal ECF, identical to the control group. Despite the lack of evidence that either anti-sporozoite or anti-merozoite antibodies are involved in immunity generated from cell line vaccination or natural infection (Irvin 1985; Hall 1988), proteins expressed by these stages have been identified as candidates for inclusion in a sub-unit vaccine. The encouraging results of preliminary vaccination trials with recombinant protein and DNA based on these antigens are discussed in detail in Section 1.10.

1.7. Epidemiology

1.7.1. Epidemiology in Tunisia

The small North African state of Tunisia has a temperate climate in the north with mild, rainy winters and hot, dry summers and is bordered by the Sahara desert in south. The principal vector of tropical theileriosis is the two-host tick *H. detritum*, as it is in other regions of the Maghreb such as Algeria and Morocco (Sergent *et al.* 1945; Flach and Ouhelli 1992; Flach *et al.* 1995). The incidence of the disease reflects the distribution and seasonal activity of this tick species with clinical cases occurring in the summer between April and September when adults are active in areas where host and vector populations co-exist. Each year around 2,500 clinical cases are recorded in Tunisia, mainly in pure-bred animals (Darghouth *et al.* 1999) in the sub-humid and semi-arid zones in the Northern part of the country (Ben Miled 1993) (see Figure 1.4.). A Tunisian Ministry of Agriculture

Figure 1.4. Bio-climatic regions in Tunisia and Turkey

(i) Tunisia

In Tunisia, the vast majority of clinical cases of tropical theileriosis occur in the sub-humid and semi-arid zones in the northern part of the country. In 1991, according to the Ministry of Agriculture, 80 % of the bovine population of Tunisia was located in the humid, sub-humid and semi-arid regions in the north of the country and approximately 85 % of clinical cases occurred in these regions (Ben Miled 1993).

(ii) Turkey

Turkey has seven distinct geographic and climatic regions ranging from the high Anatolian plateau of rolling steppe and mountain ranges to the fertile Aegean region on Turkey's western coast. Tropical theileriosis is the most important cattle disease in Turkey and has been reported from all seven geographical regions (Sayin *et al.* 2003).

Figure 1.4. Bio-climatic regions in Tunisia and Turkey

(i) Tunisia **(ii) Turkey**

study in 1991 showed that only 30 % of breeding cattle are European breeds, though they account for 80 % of registered clinical cases (Ben Miled 1993).

A female *H. detritum* may lay approximately 4,000 eggs on the ground during the summer season from which larvae hatch and search for a host, as illustrated in Figure 1.5. The larva will feed on the host and then moult to become a nymph, which will continue to feed on the same individual. Infestation of cattle with immature instars in late summer and early autumn lasts for a period of approximately 3 - 4 weeks. The engorged nymphs then detach from the host and hide in crevices of clay or stone-walled barns where they hibernate. Around April and May, when environmental conditions become move favourable, the nymphs emerge, moult into adults and begin searching for a new host. After fertilisation takes place on the host, the female will continue to feed for around a week before detaching and falling to the ground in order to lay eggs and complete the life-cycle.

A sero-epidemiological investigation of cattle herds in the semi-arid zone of Tunisia identified several states of endemicity based on the incidence and distribution of clinical cases of disease (Darghouth *et al.* 1996b). The study comprised 54 farms from the districts of Sidi Thabet and Kalaat El Andalous in Northern Tunisia. The former district was characterised as having small herds of Friesian and Holstein cattle managed semi-intensively whilst the latter consisted of larger herds of crossbred animals maintained on permanent pasture. A schizont immunofluorescence antibody test (IFAT) was used to detect circulating antibodies to *T. annulata* and calculate rates of sero-prevalence and sero-conversion. The majority of clinical cases (approximately 60 %) in the Sidi Thabet district were in adult cattle greater than three years of age, with over 90 % of farms containing animals showing positive serology. However, only one third of farms with premises in good repair had sero-positive cattle, compared with 100 % of farms that had suitable tick habitat, i.e. cracks and crevices in walls of buildings. Three states of endemicity within herds were defined –

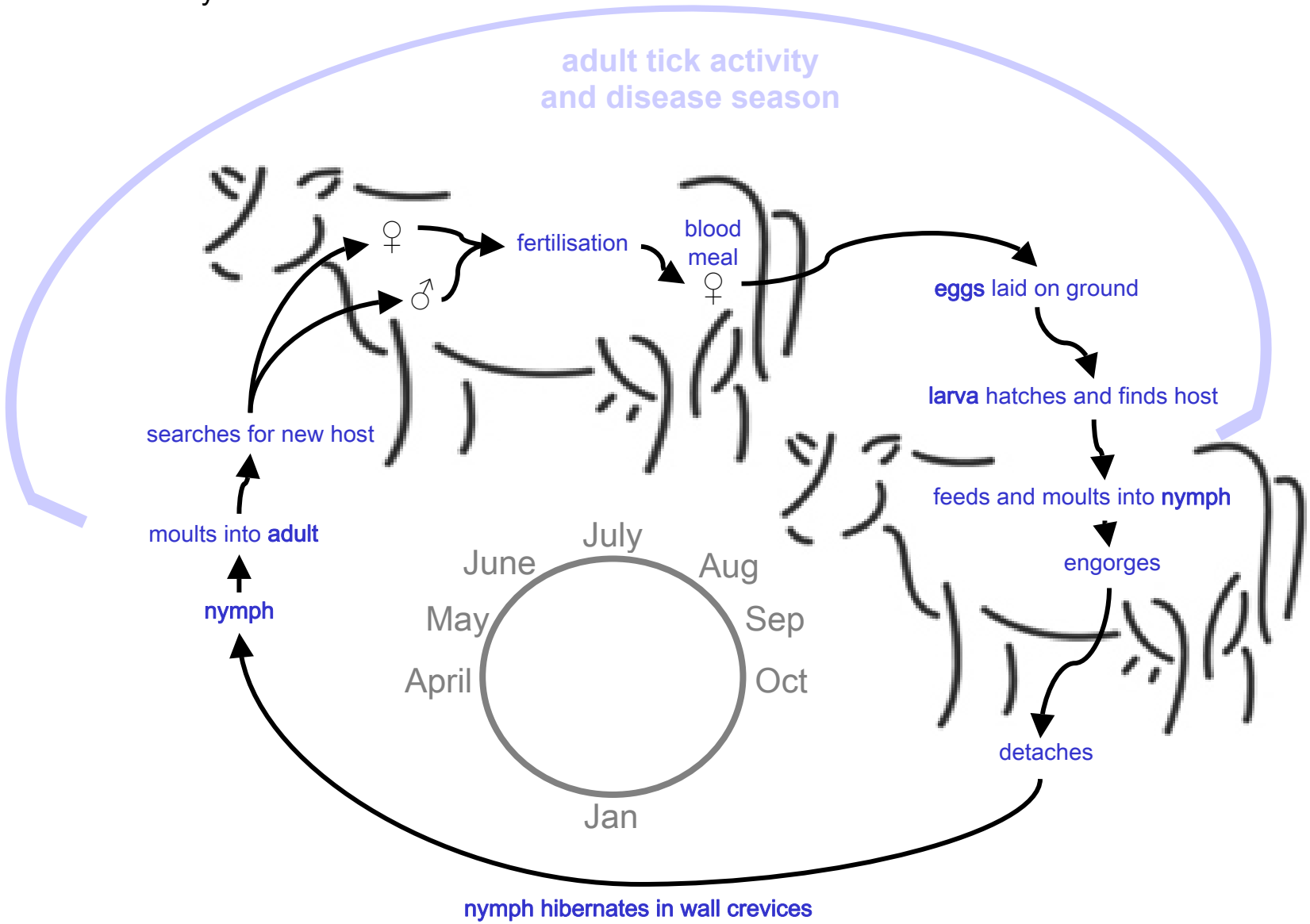
1. **Endemic stability.** Herds displayed a limited number of clinical cases in calves at the first disease season with a calf sero-prevalence of 100 % in autumn and evidence of prior exposure to a high number of ticks during summer.

2. **Low endemic instability.** Clinical cases were observed in animals exposed to less than four disease seasons, with older cattle having a sero-prevalence of 100 % and prior exposure to a moderate number of ticks.

Figure 1.5. The life-cycle of *H. detritum* in Tunisia

A female tick lays eggs on the ground during the summer season from which larvae hatch and search for a host. The larvae feed on the host before moulting to become nymphs, which continue to feed on the same individual. Engorged nymphs then detach from the host and migrate to crevices in the walls of barns where they hibernate over the winter. In spring, when environmental conditions become move favourable, the nymphs emerge, moult into adults and begin questing for a second host. After tick mating takes place on the new host, the female continues to feed for around a week before detaching and falling to the ground in order to lay eggs and complete the life-cycle. A similar vector life-cycle occurs in the Aegean area of Western Turkey, due to similar climatic conditions.

Figure 1.5. The life-cycle of *H. detritum* in Tunisia



3. **High endemic instability.** Cattle over four seasons became clinically infected while sero-prevalence was below 100 % and low tick numbers were identified during the summer months.

A study performed around the same time demonstrated that the infestation level of cattle by *H. detritum* and the prevalence of *T. annulata* infection varied between farms (Bouattour *et al.* 1996). The overall prevalence of infection of ticks was calculated at 12.4 % with females being over-represented, suggesting they may play a more significant role in disease transmission.

1.7.2. Epidemiology in Turkey

Turkey extends from 36° to 42° N and from 26° to 45° E. It is roughly rectangular in shape, is a massive 1,660 kilometres wide and is considered a transcontinental country bridging Europe and Asia. Turkey has seven distinct geographic and climatic regions ranging from the high Anatolian plateau of rolling steppe and mountain ranges to the Aegean region centred on Izmir as shown in Figure 1.4. The latter area has been described as a breadbasket, with low hills and higher mountains framing fertile valleys full of rich alluvial soil. Tropical theileriosis is the most important cattle disease in Turkey and has been reported from all seven geographical regions (Sayin *et al.* 2003). Several epidemiological studies have been conducted in the Eastern and Central regions of this large country, which is five times the size of Tunisia.

In contrast to the situation in North Africa, four species of ticks have been implicated in the transmission of tropical theileriosis in Turkey – *Hyalomma anatolicum anatolicum*, *Hyalomma anatolicum excavatum*, *Hyalomma detritum* and *Hyalomma marginatum marginatum* (Aktas *et al.* 2004). With the exception of *H. detritum*, these are all three-host ticks. Like Tunisia, disease occurs in Turkey between May and September, with the highest number of cases occurring in mid-summer, in parallel with the increase of adult *Hyalomma* ticks (Sayin *et al.* 2003).

A study in Central Anatolia focused on twelve villages around Ankara (Sayin *et al.* 1991; Sayin *et al.* 2003) examining farms with small herds of cattle which were not subject to acaricide treatment or vaccination. Similar to the situation around Aydın in the Aegean region (T. Karagenc, personal communication), many herds are moved to common pasture during the day, between March and December, before returning to the village for milking. An initial *T. annulata* sero-prevalence of 10.6 % was calculated before first disease season with 22 % of apparently naïve animals sero-converting over the course of three disease

seasons, although the number of cattle with piroplasmosis detected by blood smears was 60 %. Assuming there was no technical error in the assay used to assess sero-prevalence, this implies a separate species of bovine *Theileria* is present. It has been suggested that the second species is a member of the *T. buffeli / orientalis* group, which may be transmitted by *Haemaphysalis* ticks known to be in the area. Ticks of all four species were collected from the animals and the environment, the most heavily infected species was found to be *H. detritum*, which had an infection rate of 5 %, with a mean of 31.6 infected acini per infected tick. The study highlighted the importance of carrier animals in transmitting disease. Although the epidemiology may have been expected to be more complex due to additional tick species, the life-cycle of host and parasite in Turkey appeared similar to that in Tunisia. In the Turkish study, a single case of the disease requiring treatment was seen over the three-year period implying that there is general endemic stability in the area, in parallel with the situation in Tunisia. However, in contrast to Tunisia, local cattle breeds suffered a greater level of tick infestation and evidence of piroplasmosis, although this may have been related to differences in the management strategies for milking breeds, reducing exposure to ticks in the Turkish study. This study was unrepresentative of the area since imported cattle are understood to be at a higher risk of disease as over 90 % of animals presenting with tropical theileriosis at Ankara Veterinary Faculty clinics in 1990 were European breeds (Sayin *et al.* 1991). The number of adults infected with *T. annulata* was much greater than the number of calves infected and this is associated with heavier tick infestation in older animals.

An extensive study in Eastern Turkey collected ticks from shelters and directly from cattle (Aktas *et al.* 2004). All four *Hyalomma* species were collected from cattle, but only *H. a. anatolicum* were collected from shelters. Ticks were dissected and stained with Methyl-Green/Pyronin (Walker *et al.* 1979) to quantify the level of infected acini. 47 % of *H. a. anatolicum* collected from shelters and 19 % collected from cattle were positive for *T. annulata* infection. Similar to the Anatolian study, prevalence of infection in *H. detritum* was found to be around 5 %. As found for Tunisia and India (Sangwan *et al.* 1989), a significantly increased prevalence and intensity of infection in female ticks was observed. This arid area of East Turkey may be less hospitable to *H. detritum* than *H. a. anatolicum* and *H. a. excavatum* as *H. detritum* is normally considered a species of semi-arid areas (Bouattour *et al.* 1996). Further east, in India where the main vector is the three-host tick *H. a. anatolicum*, disease in calves represents the majority of clinical cases since they are preferentially infested with infected nymphs which may also transmit infection (Flach *et al.* 1995; Bouattour *et al.* 1996). Widely varying levels of sero-

positivity, up to 81 %, were determined in animals from eleven towns in Eastern Turkey by IFAT (Dumanli *et al.* 2005). In that study, an overall prevalence of infection of 38 % was found using the polymerase chain reaction (PCR) to amplify the TaMS1 antigen (d'Oliveira *et al.* 1995). Thus, the epidemiology in Turkey, while studied largely on the basis of prevalence and serology, appears complex with significant regional variation.

1.8. Diagnosis and treatment of tropical theileriosis

The non-specific clinical signs of pyrexia, anaemia, icterus and lymphadenopathy may be sufficient for a presumptive diagnosis of tropical theileriosis in endemic regions. This may be confirmed in the laboratory by demonstration of piroplasms in the red cells or in stained smears taken from lymph node or spleen biopsies. However, the presence of macroschizont infected leucocytes in the peripheral blood is rare and if encountered suggests a poor prognosis (Pipano and Shkap 2006). Piroplasm-infected erythrocytes may persist for years in recovered animals, whereas presence of macroschizonts indicates acute current infection. Differentiating *T. parva* from *T. annulata* morphologically is difficult, however, as previously discussed the two parasite species are, generally, clearly separated geographically. At post-mortem examination, the mucosae of the small intestine and abomasum show characteristic ulcers, 2 – 12 mm in diameter surrounded by inflammation (Kaufmann 1996), and the spleen and liver are enlarged. Impression smears of the liver, spleen and lymph nodes should be examined for macroschizonts.

The most commonly used serological test is the indirect fluorescent antibody test (IFAT). In a study in Tunisia, this technique was compared with Giemsa-stained blood smears taken from experimentally and naturally infected cattle (Darghouth *et al.* 1996c). The IFAT was able to identify more cattle exposed to *T. annulata* than parasitologically positive cases identified by blood smears, with the schizont IFAT being more sensitive than the piroplasm IFAT. An Enzyme-linked immunosorbent assay (ELISA) was initially developed to anti-*Theileria* antibodies generated from piroplasm antigens (Gray *et al.* 1980). Modern ELISAs have been developed using recombinant proteins based on the surface molecules TaMS1 (Gubbels *et al.* 2000a) and more recently the macroschizont surface protein TaSP (Bakheit *et al.* 2004; Salih *et al.* 2005).

Buparvaquone is currently the most effective drug for treating cattle suffering either East Coast Fever or tropical theileriosis (McHardy *et al.* 1985; Singh *et al.* 1993). A single treatment with buparvaquone, either at 5 mg kg^{-1} or 2.5 mg kg^{-1} intramuscularly has been shown to rapidly eliminate schizonts and piroplasms in experimentally infected calves

(Singh *et al.* 1993). However, a second treatment may be necessary if dosing at 2.5 mg kg $^{-1}$ (Mishra *et al.* 1993). 'Butalex' is the trade name for the buparvaquone preparation manufactured by Pitman-Moore, although many generic products of dubious provenance and efficacy are available in Turkey and other countries (T. Karagenc, personal communication). There have been recent reports of failure of buparvaquone treatment in both Tunisia and Turkey, which may correspond to the development of drug resistant parasites in these areas (T. Karagenç, M. Darghouth, personal communications).

1.9. Prevention and control measures

1.9.1. Resistant breeds

Disease-resistant livestock have been proposed as a sustainable method for controlling tick-borne diseases in the developing world (Glass *et al.* 2005), and a national policy in India led to a decrease in prevalence of tropical theileriosis by reducing the proportion of exotic stock in the national herd (Glass *et al.* 2005). The greater susceptibility of 'exotic' European breeds of cattle to tropical theileriosis compared to indigenous breeds has been reported in field studies in different countries (Hashemi-Fesharki 1988; Sayin *et al.* 1991; Darghouth *et al.* 1999). When *Bos taurus* and *Bos taurus* / *Bos indicus* crossbred (Sawihal) cattle were challenged with a graded dose of *T. annulata* sporozoites, the clinical outcomes were found to be dose-dependant in crossbred but not in taurine calves, suggesting a degree of resistance in the Sawihal cross calves (Preston *et al.* 1992). A separate study using pure Sawihal calves demonstrated a significantly reduced severity of infection compared with Holstein calves (Glass *et al.* 2005). In a recent Sudanese study, indigenous Kenana and Friesian calves were experimentally infected with a lethal dose of *T. annulata* (Bakheit and Latif 2002). Only two Kenana cattle required treatment, compared to complete mortality in Friesian calves. The reduced clinical severity was related to the Kenana calves exhibiting a lower schizont parasitosis and reduced leucopaenia with less damage to lymphoid tissues. It has also been suggested that cattle resistant to the tick *Rhipicephalus appendiculatus* may be important in promoting endemic stability of the related disease East Coast Fever caused by *Theileria parva* (Fivaz *et al.* 1989), however no study on the application of tick resistant cattle to controlling tropical theileriosis has yet been undertaken. Despite clear evidence that European breeds show greater susceptibility, it should be borne in mind that they were introduced to improve the productivity of local breeds. It is unlikely that as agricultural practices advance through many parts of the developing world that there will be significant reversion and reliance on local breeds.

1.9.2. Vector control

In endemic areas, farm management and tick vector control traditionally play a significant role in reducing the impact of theileriosis. The main vector in North Africa is the endophilic *H. detritum*, which lives in cracks and crevices in the walls of cattle accommodation. Consequently, improvement of cattle housing by eliminating such areas should reduce the risk of disease transmission. Additionally, regular use of acaricide on animals and housing, especially during peak tick activity season should assist disease control. However, the problem with completely eradicating the tick population from a locality is the creation of a zone of endemic instability, where there is insufficient natural challenge to allow immunity of the herd to be maintained. A large naïve cattle population may be created which risks decimation following a breakdown in control measures. For example, with *T. parva* in Zimbabwe when cattle dipping programmes were suspended during the civil war, there was a rapid increase in both tick numbers and outbreaks of tick-borne disease (Pegram *et al.* 2000). Additionally, acaricides may contaminate milk and meat and some preparations are known to be a risk for human health. Acaricide resistance has also been reported (Musoke *et al.* 1996). Currently it is considered that acaricides are likely to be more effective when used in an integrated control strategy. An imaginative study using pyrethroid impregnated decoy ticks of another *Hyalomma* spp. demonstrated some success in attracting adult males (Abdel-Rahman *et al.* 1998) on camels though it is doubtful this can be applied to the bovine disease on a large scale.

1.9.3. Historical and current forms of immunisation

1.9.3.1. Infusion of infected blood

The first attempt at immunising cattle against tropical theileriosis was in the 1930s in Algeria. Blood donated from clinically affected animals, infected with a strain of low virulence, was subject to mechanical passage between cattle with the parasite losing its ability to differentiate into the merozoite stage (Sergent *et al.* 1945). Protection was estimated at one year, in the absence of natural challenge. In Israel in the 1940s a similar regime was used – cattle were inoculated with a Tunisian strain of low virulence, then two months later were boosted with a local strain to reinforce the immunity and presumably account for local strain variation (Pipano and Shkap 2006). Such vaccines are now considered to pose an unacceptable risk to the recipient. In addition to causing clinical tropical theileriosis there is the risk of transmission of other blood-borne pathogens. This has been demonstrated by contamination of a batch of babesiosis and anaplasmosis vaccine in Australia when screening for Bovine Leucosis Virus (BLV) failed (Rogers *et al.* 1988).

1.9.3.2. Infection and treatment

Infection and simultaneous treatment with tetracyclines has been shown to be effective against *T. annulata* (Gill *et al.* 1976). Naïve calves infected and dosed with chlortetracycline developed a mild form of the disease and were found to be solidly resistant to subsequent severe homologous challenge. This infection-treatment technique was pioneered for East Coast Fever by Neitz in the early 1950s and is still in effective usage today (Kanhai *et al.* 1997). A crude suspension of *T. parva* sporozoites is prepared from adult ticks fed on rabbits, in order to allow the sporozoites to mature, and the suspension mixed with glycerol and stored in liquid nitrogen. An aliquot of this mixture is thawed and then injected subcutaneously concurrently with an intramuscular injection of long-acting oxytetracycline. Following extensive cross-immunity trials, the 'Muguga cocktail' was developed to protect against most strains present in East Africa and is known to contain at least three genotypes of *T. parva* (Bishop *et al.* 2001). While generally effective, a recent molecular study provided evidence of breakthrough infection (Oura *et al.* 2004). When the main *T. annulata* chemotherapeutic, buparvaquone, was used in place of tetracycline in the *T. parva* infection and treatment protocol, it was found to be too effective, with infection failing to develop and the recipient remaining partially or fully susceptible (Ngumi *et al.* 1992). Despite the recorded effectiveness of infection and treatment, generally the technique is considered relatively expensive and may pose considerable logistical problems since the maintenance of a cold-chain is required. Additionally, incorrect administration of immunogen or drug, undiagnosed intercurrent disease or poor nutrition status of the recipient may precipitate clinical cases. Furthermore, it may not protect against all strains present in a locality as well as potentially introducing the 'vaccine' strains into the field population.

1.9.3.3. Cell line vaccines

The capability of inoculated macroschizonts to invade and establish themselves in mononuclear lineages is the basis for the various cell line vaccines currently deployed in endemic regions around the world. Similar to the serial bovine passage system for creating blood vaccines, schizont infected cell line cultures are passaged *in vitro* to achieve attenuation of virulence (Darghouth *et al.* 1996a; Pipano and Shkap 2000). The ability to maintain such cultures indefinitely has been the cornerstone of vaccine production since the discovery of the phenomenon of attenuation in the mid 1960's (Tsur and Pipano 1966). Attenuation has been related to the reduction in the number of genotypes, with optimal clinical results demonstrated against homologous challenge (Gill *et al.* 1980; Darghouth *et*

al. 1996a) and good resistance to heterologous challenge also reported (Hashemi-Fesharki 1988; Darghouth *et al.* 1996a). Both susceptible indigenous and exotic cattle are effectively protected.

Different vaccination protocols have been implemented with varying degrees of success in several countries, making comparisons on safety and efficacy difficult. Cultured macroschizont-infected cell lines may be cryo-preserved in liquid nitrogen and between 5×10^5 and 5×10^6 cells are administered after three months to three years of *in vitro* cultivation. In China between 1975 and 1990, 1.8 million cattle were vaccinated, with a protection rate of 99.58 % (Zhang 1991) without any detectable reduction in milk production or increased incidence of abortion. An Iranian study found no significant changes in haematological parameters apart from a mild leucocytosis which subsided three weeks post inoculation (Hashemi-Fesharki 1991). The duration of immunity is reported to last beyond three and a half years (Zablotskii 1991) although some animals are reported to suffer severe clinical theileriosis four or more years post vaccination (Pipano and Shkap 2006). This suggests animals may require re-vaccination should there be insufficient natural challenge. However, following an allogeneic response to previously inoculated donor-leucocytes, circulating antibodies may block transfer of parasite to the recipient, on subsequent use of the same cell line vaccine. Fortunately, use of a second vaccine strain (with presumably a different host MHC phenotype) considerably enhances immunity (Pipano and Shkap 2006).

It has been shown that it is unnecessary to define the Bovine Leucocyte Antigen (BoLA) profile of the donor cells and recipient cattle before primary immunisation (Innes *et al.* 1989b), as transfer of *T. annulata* parasites from the cell line vaccine to the recipient appears to be independent of histocompatibility. This is in marked contrast to the situation in *T. parva* where MHC class I genotypes of donor and recipient must be matched before transfer of the parasite occurs (Dolan *et al.* 1984). Attenuation of *T. parva* infected cell lines *in vitro* has also been shown to be accompanied by a loss of immunogenicity (Brown 1981). These differences account for the reliance on the infection and treatment form of vaccination to control East Coast Fever and cell line vaccination to control tropical theileriosis.

Altered gene expression accompanying selection occurring in early passage of virulent cell lines has been suggested (Sutherland *et al.* 1996), however the molecular mechanisms for attenuation have not yet been elucidated. It has been proposed that a loss of matrix metalloproteinases activity may account for loss of clinical features such as metastasis,

abomasal ulceration and cachexia (Adamson and Hall 1996; Adamson and Hall 1997; Adamson *et al.* 2000). It is of interest that parasite proteins transported to the host nucleus could influence the gene expression profile of attenuated cell lines. An increase in expression of the parasite encoded host-nuclear protein, TashHN, has been documented in attenuated lines (Swan *et al.* 2003), although it is unclear how this may contribute to a reduction in virulence. A study on a Turkish vaccine cell line highlighted the reduced ability to differentiate towards the merozoite stage and using mRNA differential display, demonstrated differing gene expression profiles in attenuated and non-attenuated cell lines (Somerville *et al.* 1998). However, as virulent field strains which do not produce significant piroplasmosis exist, inability to differentiate to this stage is in itself unlikely to be a major cause of attenuation (Somerville *et al.* 1998). Cell line immunisation does not necessarily prevent piroplasmosis and the induction of carrier status (Zablotskii 1991). Such piroplasms may be the result of super-infection following challenge by feeding ticks or differentiation of a limited number of macroschizonts derived from the vaccine inoculum.

Although current cell line vaccines are relatively inexpensive to produce, their use is curtailed due to the expense of maintaining a cold chain and the technical expertise in handling and administering the vaccine correctly. This has been overcome, to a degree, in Tunisia where the vaccine is thawed and refrigerated at 4 °C for up to one month prior to use (M. Darghouth, personal communication). The Chinese also report storing vaccine for up to two months at 4 °C without detriment (Zhang 1991). A further disadvantage is that tissue culture vaccines can pose a risk to lactating or pregnant cows due to transient pyrexia post-vaccination (Hashemi-Fesharki 1988). Infectivity of blood-derived vaccines have to be assayed by titration in susceptible cattle (Pipano 1997) although this does not guarantee there will be no future reversion to virulence. Therefore, despite the success of attenuated culture vaccines, there is still a need for development of an effective sub-unit vaccine that could remove some of the inherent problems of the live vaccines.

1.10. Sub-unit vaccines

The various safety and efficacy problems discussed above have redirected research towards developing a sub-unit vaccine comprising specific parasite antigens, intended to stimulate a protective immune response. Such a vaccine will also have the distinct advantages of stability (obviating the need for cold chains) and having a defined composition (allowing for simple quality control). Identifying genes encoding these protective antigens is the focus of current research. Once identified, recombinant proteins

or DNA based on the relevant genes can be produced and considered for inclusion in a sub-unit vaccine. Three potential stages of the *T. annulata* life-cycle have been suggested as targets: the sporozoite, in order to reduce the infective dose; the macroschizont, in order to control proliferation of mononuclear cells and pathology and the merozoite / piroplasm, in order to reduce infection of erythrocytes and anaemia, and to limit disease transmission to the vector.

As with any other disease, account must be taken of the diversity of antigens in the natural population, to which the vaccine is designed. Antigens already characterised at each stage are discussed in turn below.

1.10.1. Sporozoite antigens

Research initially focused on the sporozoite stage using a similar approach to that taken in East Coast Fever (Williamson *et al.* 1989), as it is known that cattle repeatedly exposed to live sporozoites develop a humoral response to that stage (Williamson 1991). Following experimental infection with inactivated sporozoites, challenged cattle were able to retard development of macroschizonts, however it is unlikely that anti-sporozoite immunity alone would protect against tropical theileriosis in the field (Williamson 1991). Work in *T. parva* suggests free sporozoites may exist for less than ten minutes in the circulation (Fawcett *et al.* 1982), providing a brief window of opportunity for the immune system to detect and neutralise every one. It may, however, reduce the infective dose sufficiently to allow a protective immune response to develop to subsequent stages.

A gene product specifically targeted by the humoral immune response to the *T. annulata* sporozoite was identified in 1989 and subsequently named SPAG-1 (Williamson *et al.* 1989). A series of hybridoma cultures secreting anti-sporozoite monoclonal antibodies (mAbs) were identified using IFAT on fixed sporozoites. An *in vitro* assay was used to screen these mAbs for their ability to block sporozoite invasion of mononuclear cells. mAb 1A7 was found to be highly effective in blocking invasion and was subsequently used to screen a *T. annulata* λgt11 genomic expression library to identify clones carrying the epitope. The clone (λgt11-SR1) was identified and used to obtain a full-length genomic clone representing an allele of the Hissar strain. This was later fully sequenced and found to code for a 91.9 kDa polypeptide of 907 amino acids (Hall *et al.* 1992). Repetitive regions containing multiple copies of the peptide motifs, `PGVGV` and `VGVAPG`, were identified towards the N-terminus of the gene. These motifs had a high degree of homology with the bovine elastin repeat at the amino acid level, although the nucleotide

sequence was more divergent. At the N-terminus, 18 mainly hydrophobic amino acids were identified as a signal sequence, implying export from the parasite. Furthermore, a membrane anchor was suggested at the C-terminus with the presence of a 24-residue hydrophobic domain. Two explanations for the mimicry of the bovine-elastin domain were proposed. First, the parasite may be mimicking the host in order to evade the immune response. Ironically, the high level of homology to bovine elastin raised concerns that if the full-length SPAG-1 was used as a sub-unit vaccine either an autoimmune disease would be precipitated or the host would not recognise the immunogen as foreign and fail to mount a response. Alternatively, a functional molecule may be mimicked in order to assist the parasite in some activity. The authors speculate that the SPAG-1 antigen functions as a ligand for host cell recognition. The elastin receptor is known to bind the VGVAPG elastin peptide (Blood *et al.* 1988; Robert *et al.* 1989; Mecham *et al.* 1989), which exhibits positive chemotaxis towards the monocytic cell type that *T. annulata* preferentially invades. Furthermore, it was found that there is a high level of synonymous substitutions within elastin repeats of available sequences when compared to the bovine sequence suggesting purifying selection may be acting on these areas to conserve the amino acid sequence.

Katzer *et al.* (Katzer *et al.* 1994) showed that this sporozoite surface antigen is encoded by a single copy gene and demonstrated a high degree of polymorphism by comparing two full-length alleles. An overall polypeptide identity of 92 % was observed with the N- and C-termini containing the most conserved areas. The most variable region was found in the middle of the molecule where amino acid identity fell to 60 %. One of the SPAG-1 alleles did not contain any VGVAPG motifs, casting doubt on both the elastin ligand and self-mimicking immune evasion hypotheses. The C-terminal half of SPAG-1 was identified as the most conserved region and it was suggested this may be of use when designing a sub-unit vaccine. A companion study located neutralising determinants at the C-terminus of the molecule using monoclonal antibody (mAb) 1A7 (Boulter *et al.* 1994). Restriction fragment length polymorphism (RFLP) analysis, which was used to identify the two full-length sequences, suggested a further two allelic variants (Katzer *et al.* 1994). Furthermore, this region of the molecule showed 56 % identity compared to the orthologous p67 protein of *T. parva*, implying that a common vaccine against the two parasites may be feasible.

Following this study, a C-terminal fragment of SPAG-1 was expressed as a fusion protein with the hepatitis B core antigen (Boulter *et al.* 1995). Cattle immunised with this showed

a high titre of invasion-neutralising antibody with IFATs confirming that the antibodies recognised the surface of the sporozoite. The study also indicated that native SPAG-1 is capable of inducing a good T-cell response and therefore it must also contain T-cell as well as B-cell epitopes. Encouragingly, the vaccinated cattle showed some attenuation of clinical signs, compared to the control group, with an extended time before macroschizonts appeared in blood smears.

In order to optimise the partial protection observed after challenge, four different delivery systems and adjuvants were compared (Boulter *et al.* 1999), including two recombinant p67 formulations. Unfortunately, none of the regimens tested offered complete protection, although a positive outcome was recorded with three of the four. SPAG-1 administered with the saponin-based adjuvant SKBA (Smith Kline Beecham Adjuvant) was the most effective and a degree of cross-protection with the p67 vaccine was noted in this trial.

An ECF vaccine trial using recombinant p67 antigen fused with NS1 protein of influenza virus A has shown encouraging results (Musoke *et al.* 1992). Homologous challenge with a *T. parva* sporozoite stabilate induced anti-sporozoite neutralising antibodies in cattle and resulted in protection of six out of nine calves from severe clinical disease. In contrast, all the control animals succumbed to severe ECF, with the majority being euthanased *in extremis*. Although the nature of the induced protective immune response was unclear, limited establishment of sporozoites appears to play a role since a paucity of macroschizonts was noted in many of the trial group. Several mechanisms of action have been proposed: (a) opsonisation of sporozoites, thereby enhancing phagocytosis; (b) antibody dependent cell-mediated cytotoxicity (ADCC) and (c) a contribution of complement to neutralise sporozoite infectivity.

Conservation of neutralising determinants of the sporozoite antigens between the two species stimulated a cross-reactivity study using mAbs and sera raised against each antigen (Knight *et al.* 1996). Cross-reactive linear epitopes were identified in the conserved N- and C-termini using mAb 1A7. It was speculated that host cell recognition determinants are outside these regions due to the difference in cell tropism between the parasite species. In addition to determinants at the C-terminus, epitope mapping suggests that there may be important determinants throughout the molecule, including the variable regions. A recent field study using a sub-unit vaccine based on p67 (Musoke *et al.* 2005) compared a near full-length sequence with an 80 amino acid C-terminal fragment, following experimental evidence of the efficacy of the latter (Bishop *et al.* 2003). Groups of cattle were immunised with one or other of the constructs and exposed to natural tick challenge. The

incidence of severe East Coast Fever was reduced and there was no difference in the performance of the two vaccines.

The *SPAG-1* orthologue in *T. lestoquardi* was recently identified and imaginatively named *SLAG-1* (Skilton *et al.* 2000). The gene is predicted to encode a 76 kDa protein of 723 amino acids, which is comparable to p67, but shorter than SPAG-1. The highest sequence identity between the *T. annulata* and *T. lestoquardi* genes is in the first 100 amino acids and the C-terminal half of the polypeptide sequence. Although inhibition of infectivity of *T. lestoquardi* sporozoites has not been demonstrated experimentally, the peptide sequence `PSLVI` that encodes the epitope recognised by mAb 1A7 is present in SLAG-1.

1.10.2. Macroschizont antigens

Cell-mediated responses to the macroschizont-infected leucocyte are considered to be responsible for natural protective immunity (Section 1.6.2.1.) and it has been suggested that schizont antigens will be an important component of any recombinant vaccine against tropical theileriosis (Preston *et al.* 1999). With natural immunity to infection being, in general, cross-protective, it would clearly be important to identify the schizont antigens responsible for eliciting this response. Moreover, pathogenesis of *T. annulata* infection is primarily attributed to the uncontrolled multiplication of infected leucocytes (Section 1.5.). Consequently, a vaccine directed against this stage of infection, which prevents or reduces host cell proliferation and metastatic behaviour, may at the very least reduce the severity of the disease.

It was demonstrated that, in *T. parva*, $CD8^+$ T-cells are capable of controlling infections when immunity was successfully transferred from an immune to a non-immune twin calf (McKeever *et al.* 1994). Cytotoxic T-lymphocytes (CTLs) are highly specific, recognising only schizont-infected cells and it has been shown that CTL responses are parasite strain specific (Morrison *et al.* 1987; Morrison 1996). As the macroschizont is intracellular and the mechanisms of protective immunity involve the recognition of the infected cell, the relevant antigens will be presented on the surface and, in the case of *T. parva*, as peptides presented on MHC Class I molecules (Morrison and McKeever 1998). While a number of schizont antigens have been detected using immune sera, these are unlikely to be involved in the protective immune response. The identification of the antigens presented on MHC Class I molecules has provided a major technical challenge in terms of their identification. As it is predicted that proteins secreted by the macroschizont enter the MHC I processing pathway, attempts have been made to identify this particular subset of molecules. In the

pre-genomic era, a 'signal sequence trapping' method, which involved functional analysis in *Saccharomyces cerevisiae*, was used to predict secreted proteins (Musembi *et al.* 2000). However, with the advent of the published genome sequence of *T. parva*, it has been possible to screen the genome *in silico* for genes encoding a secretory signal peptide (Gardner *et al.* 2005). Of the 986 predicted genes on *T. parva* chromosome I, a subset of 55 were predicted to encode secreted antigens. 36 of these genes were cloned and along with a series of random schizont cDNA clones they were used in an immuno-screening approach to identify MHC I presented antigens (Graham *et al.* 2006). This relied on screening transiently transfected antigen presenting cells with fully characterised CTL from cattle immunised with a live vaccine (Taracha *et al.* 1995); the methodology of this study is further discussed in Section 4.4.6. The six genes that encoded products recognised by cytotoxic T-cells were annotated as the ε-subunit of T complex protein 1, elongation initiation factor 1A, heat-shock protein 90 (HSP90), cysteine protease and two hypothetical proteins. Five of these proteins were suggested as vaccine candidate antigens, with the exception of HSP90, which failed to be recognised by CTLs from immune cattle resolving a challenge infection. The immunogenicity of the five candidate antigens was tested *in vivo* using a prime / boost immunisation protocol. 24 cattle were immunised with all five antigens and following booster immunisation were challenged with a lethal dose of sporozoites. Antigen-specific $CD8^+$ T-cell mediated IFN-γ responses and $CD8^+$ CTL activity was monitored throughout the experiment. Nineteen cattle showed an IFN-γ response while only seven cattle exhibited CTL activity. Survival to challenge was found to be associated with cattle exhibiting a CTL response ($p < 0.001$; χ^2 test, comparing vaccinated responders to vaccinated non-responders). Unvaccinated control animals, which were subjected to challenge, all developed severe East Coast Fever. Although encouraging, it should be noted that the majority of cattle vaccinated with all five antigens also developed severe clinical signs of East Coast fever and needed to be euthanased. Unfortunately, in *T. annulata*, no equivalent CTL lines have been reported, thus currently precluding this approach.

One of the few macroschizont antigens that has been characterised in detail is the polymorphic immunodominant molecule of *T. parva*, PIM. This protein was identified using a panel of monoclonal antibodies and subsequently localised by immuno-electron microscopy to the surface of the macroschizont (Shapiro *et al.* 1987). The size of this antigen ranged between 68 kDa and 95 kDa in different parasite isolates, with expression later confirmed in the sporozoite stage (Toye *et al.* 1991). In addition to recognition by mAbs, antisera from immune cattle recognised this antigen in schizont-infected

lymphocytes on Western blot analysis (Toye *et al.* 1991), with differential expression of antigenic epitopes observed between *T. parva* stocks. cDNA sequences from two *T. parva* stocks suggested a 5′ conserved region of 71 amino acids and a 3′ conserved region of 208 amino acids, bridged by a central variable region where amino acid identity fell to 50 % (Toye *et al.* 1995b). Several types of repeated motifs were identified in the central region, the most distinctive of which is the tetra-peptide `QPEP`, which is tandemly repeated. Similar to p67, anti-PIM mAb is able to inhibit sporozoite infectivity *in vitro* (Toye *et al.* 1995b).

Probing Southern blots of genomic DNA with a Muguga *PIM* cDNA, confirmed that *PIM* is single-copy gene (Toye *et al.* 1995a), with size polymorphism observed in a panel of genomic DNA representing *T. parva* stocks from Eastern and Southern Africa. When 5′ sequences of buffalo-derived *PIM* were compared against the Muguga sequence a high level of non-synonymous substitutions was identified. Non-synonymous substitutions are defined as single nucleotide changes in DNA sequence, which encode a variant amino acid. In contrast, synonymous or silent substitutions do not result in amino acid change. The rate of non-synonymous (d_N) to synonymous substitutions (d_S), often referred to as $d_N d_S$ is a useful index for quantifying the influence of purifying and diversifying selection. In this particular study, the ratio of synonymous substitutions per synonymous site to that of non-synonymous substitutions per non-synonymous site was calculated using Nei's method (Nei and Gojobori 1986). Pair-wise comparisons between the Muguga and two buffalo-derived *PIM* sequences suggested selection for diversification was operating on the *PIM* gene. The 3′ region of the Muguga *PIM* gene contains two introns at 55 and 61 bp (Toye *et al.* 1995b), and in contrast to the diversity observed in the coding regions, both introns were found to be highly conserved among several cattle-derived and buffalo-derived sequences. Taken together, this provides evidence that the protein is undergoing rapid diversification over parts of its coding region, perhaps as the result of selective pressure from the bovine immune system (Toye *et al.* 1995a).

More recently, the *PIM* orthologue in the genome of *T. annulata* was identified and named *TaSP* (*Theileria annulata* Surface Protein) (Schnittger *et al.* 2002). Similar to *PIM*, *TaSP* is a single-copy gene and is expressed by the macroschizont and sporozoite stages. However, the gene is predicted to encode a smaller protein of 315 amino acids with a molecular weight of only 36 kDa. A global amino acid identity of 55 % between TaSP and PIM was calculated, ranging from 93 % at the termini to 10 % in the central region; and analogous to PIM, the central region of TaSP contained an abundance of negatively

charged amino acids. A signal sequence of 19 amino acids was identified at the N-terminus, while three transmembrane domains were found at the C-terminus, suggesting the central and N-terminal regions have an extracellular location.

To investigate allelic diversity, several clones representing widely separated geographical isolates were sequenced. Diversity was observed not only in sequences from different parts of the world, but also in clones derived from a single isolate. The first 37 and last 121 amino acids were found to be strongly conserved, while the central portion (position 154 – 171) showed amino acid diversity and length polymorphism. Although this general structure is similar to PIM, the polymorphic central region of TaSP is much shorter, the allelic polymorphism is lower and the characteristic repetitive motifs are absent. The authors speculated that the diversity is a result of mutation and intragenic recombination.

A recombinant TaSP polypeptide, corresponding to amino acids 26 – 157, was used to generate a specific antiserum. An IFAT against macroschizont-infected cells was performed with this antiserum resulting in staining of the surface of the macroshizont and the parasite nuclei. Eight different antisera from *T. annulata* infected cattle all reacted strongly with the recombinant polypeptide and support the view that TaSP, like PIM, may be an immunodominant molecule. No cross-reactivity studies have been undertaken between *T. parva* and *T. annulata* and it is unknown if mAbs and sera raised against TaSP react against *T. parva* macroschizonts or vice versa. However, in a recent study, sera from small ruminants infected by a *Theileria* species in China were probed for reactivity with TaSP (Miranda *et al.* 2004). The majority of sera reacted to this protein in both ELISA and Western blots, demonstrating cross-reactivity with TaSP and suggesting a high level of conservation of this antigen between these species.

Bioinformatic screening of TaSP identified 22 epitopes predicted to bind to the bovine MHC class I molecule (A20), however only six were located in the polymorphic central region (Schnittger *et al.* 2002). Class I restricted cells may present the molecule for recognition by $CD8^+$ cytotoxic T-cells. The prediction of 14 of these epitopes in conserved areas suggests they may act as cross-protective determinants, conferring immunity to heterologous challenge following immunisation. In addition to predicted T-cell epitopes, both PIM and TaSP clearly present B-cell epitopes that may be recognised by the immune response. A prior epitope-mapping study of PIM demonstrated at least ten different B-cell epitopes throughout the molecule (Toye *et al.* 1996). Bovine antisera reacted strongly with the tetra-peptide in the polymorphic central region of PIM. Although the bovine antisera were unable to neutralise sporozoite infectivity *in vitro*, mouse-derived mAbs, which did

not react with the tetra-peptide repeat, were able to neutralise *T. parva* sporozoites, *in vitro*. This result demonstrated the inability of bovine antiserum to inhibit sporozoite infectivity, however the cattle did produce antisera that reacted to the same peptides as the neutralising murine mAbs, suggesting cattle may produce neutralising antisera if immunised with PIM.

1.10.3. Merozoite / piroplasm antigens

One of the most intensively studied molecules in *T. annulata* is the immunodominant merozoite and piroplasm antigen TaMS1, following its discovery on the surface of merozoites derived from the Ankara (A_2) cell line (Glascodine *et al.* 1990). Subsequently, two variants with molecular masses of 30 kDa and 32 kDa were identified from lysates of infected cloned cell lines derived from the *Ta*A_2 stock (Dickson and Shiels 1993). When piroplasm extracts of several *T. annulata* stocks representing different geographical regions were compared, stocks could be classed according to the 30 kDa / 32 kDa profiles. In addition to finding antigenic epitopes conserved between the alleles, it was demonstrated that unique epitopes exist on both proteins (Dickson and Shiels 1993). Further studies on *T. sergenti* resulted in characterisation of the gene encoding the major merozoite / piroplasm surface antigen (MPSA) (Sugimoto *et al.* 1991). Antigenic diversity as a result of coding for variant amino acid sequences was later demonstrated for this 32 kDa protein (Kubota *et al.* 1996), along with a smaller 23 kDa merozoite/piroplasm protein (Zhuang *et al.* 1995).

The potential of TaMS1 to be a component in a sub-unit vaccine or to be employed as a diagnostic reagent prompted Shiels *et al.* to investigate the generation of diversity in detail at the nucleic acid and amino acid level (Shiels *et al.* 1995). A restriction fragment length polymorphism study (RFLP) was undertaken, which successfully differentiated the variant fragments representing the two previously described antigenic forms and suggested an even greater level of diversity. The analysis also indicated that the *TaMS1* gene is single-copy, hence the variant RFLP profiles correspond to different allelic forms of the gene in distinct haploid parasite genotypes. The gene was named *TaMS1* and the alleles representing the two previously described size-variants named *TaMS1-1* (30 kDa) and *TaMS1-2* (32 kDa). Both polypeptide variants comprise 281 amino acids, with a signal sequence at the N-terminus and a putative GPI-anchor motif at the C-terminus. Sequences of orthologous genes in the *T. buffeli / orientalis* group, *T. sergenti* and *T. parva* (*TpMS1*) were available or obtained and compared with *TaMS1-1* and *TaMS1-2* sequences. A high level of polymorphism at the amino acid level was observed between residues 50 and 60 where N-linked glycosylation sites were predicted. Thus, it was postulated that diversity

of glycosylation pattern could account for the difference in molecular weight between TaMS1-1 and TaMS1-2 polypeptides. This was later discounted by Katzer *et al.*, who were unable to find evidence that the molecules underwent N-linked glycosylation (Katzer *et al.* 2002). In this study epitope mapping was also undertaken using an array of overlapping expressed fragments of TaMS1. The results revealed that a mAb and antisera raised against the native protein reacted against divergent epitopes that were dependent on a tertiary molecular conformation sensitive to oxidation by periodate. The results also showed that conserved epitopes could be masked by the tertiary conformation of the molecule, thus limiting exposure a protective bovine immune response. These variable domains are more likely to be exposed, while conserved domains may be hidden. It was also shown that TaMS1 has the potential to form complexes with itself, perhaps forming a polymeric lattice on the surface of the parasite. One may speculate that this quaternary structure plays a role in forming a protective coat around the surface of the merozoite and piroplasm. The presence of TaMS1 on the surface on the merozoite has been demonstrated unequivocally by immuno-electron microscopy (Glascodine *et al.* 1990).

In order to determine whether divergent allelic forms could be categorised according to place of origin, five alleles of *TaMS1* were sequenced, representing three different geographical regions (Shiels *et al.* 1995). No correlation was observed, with similar sequences coming from different areas and divergent sequences representing the same area, suggesting a rapid spread of novel genotypes across large distances. Furthermore, Southern blotting suggested a high level of heterogeneity within the Tunisian population. To further investigate this diversity, an extensive allele sequencing project was undertaken (Katzer *et al.* 1998; Gubbels *et al.* 2000b). 129 *TaMS1* sequences were obtained for parasites isolated in North Africa, Southern Europe, the Middle East and India to examine diversity over large geographical distances (Gubbels *et al.* 2000b). At least one *TaMS1* sequence was obtained from each isolate, with many isolates providing multiple allelic sequences. At both nucleotide and amino acid level, almost the same amount of sequence divergence was demonstrated within *TaMS1* alleles, as between *TaMS1* and the orthologous *TpMS1* gene sequence, suggesting an active process may be driving amino acid diversity of *TaMS1* antigens within *T. annulata*. More formal evidence for a process of positive selection was provided when $d_N d_S$ analysis was undertaken. A sliding windows method, examining a 20-codon frame, was performed on 18 representative sequences selected from the complete dataset. Seven polymorphic polypeptide regions distributed over the entire length of the molecule were suggested as being positively selected.

No correlation between geographical origin and sequence similarity could be made, with virtually identical alleles found in Tunisia, Mauritania and India, confirming the findings of the previous more limited studies. Heterogeneity, within Tunisia itself, was comparable to that observed over the entire dataset. Consequently, cluster analysis failed to group sequences derived from the same region, resulting in a random distribution on a similarity dendrogram.

When line-ups based on amino acid sequence were performed, conserved regions were found interspersed in the variable central portion of the molecule including the conserved motifs of KE (Lys-Glu) and KEL (Lys-Glu-Leu) that have been implicated in erythrocyte invasion (Molano *et al.* 1992).

In order to find evidence for intragenic recombination, the computer software PLATO (Grassly and Holmes 1997) was utilised. Using a maximum likelihood (ML) method, this program identified six regions where intragenic recombination is predicted, including two putative glycosylation sites. Thus, the molecule appears to have a mosaic-like structure comprised of blocks of variant sequence, which may be shuffled between alleles during intragenic recombination. A similar mosaic structure was subsequently described in *PIM* in *T. parva* (Geysen *et al.* 2004) and it is noteworthy that the major immunodominant surface polypeptides of two different stages are divergent and have a similar basic structural composition. It was suggested that conserved sequences have a functional / structural role, while the variable regions are under evolutionary pressure to diversify in order to evade the bovine immune system (Gubbels *et al.* 2000b). Therefore, the molecule may be recognised by a protective immune response. A monoclonal antibody raised to the orthologue of TaMS1 in *T. sergenti* has been shown to inhibit merozoite invasion of erythrocytes *in vivo* after transfusion into a *T. sergenti* naïve calf (Tanaka *et al.* 1990).

Recombinant proteins representing the internal sequence of both TaMS1-1 and TaMS1-2 have been expressed in *E. coli* and *S. typhimurium* (d'Oliveira *et al.* 1996) for the purpose of trial immunisation of cattle. All recombinants were recognised by immune calf serum prompting optimism about the usefulness of the antigen as a vaccine candidate. Subsequently, a variety of delivery systems were tested in the first vaccine trial based on TaMS1 (d'Oliveira *et al.* 1997). The most promising results were obtained by immunising with (i) recombinant protein presented with an immuno-stimulating complex (ISCOM) and (ii) naked plasmid DNA. Four weeks after the last immunisation, all the calves were subjected to challenge from a *T. annulata* blood stabilate. All the calves immunised with recombinant protein did not develop clinical signs of tropical theileriosis and were

protected from disease. The level of parasitaemia and the reduction in packed cell volume (PCV) was dramatically lower in this group compared with unimmunised control animals. Two of the three calves immunised with naked DNA were also protected against disease and the other developed a non-fatal clinical disease and recovered without any treatment. The three ISCOM immunised calves all generated sera that recognised the surface of the piroplasm by IFAT before the calves were challenged, suggesting that the merozoite surface may also be recognised. Protection may therefore be due to high levels of circulating anti-TaMS1 antibodies binding to the merozoite surface and either inhibiting invasion of red cells or opsonising the parasite. However, due to the nature of the challenge method, the study assumes that clinical signs were due to the erythrocytic phase of the disease alone. Interestingly, the two calves protected by the DNA vaccine generated no detectable antibodies against the recombinant proteins, suggesting that this immunity is T-cell dependent. These encouraging preliminary results demonstrate that TaMS1 can generate a protective immune response in cattle, although it is critically influenced by the method of delivery and the nature of the protective response has still to be defined. Additionally, TaMS1 has been utilised in an attempt to improve the capacity of the *T. annulata* sporozoite antigen SPAG-1 to elicit a protective response (Boulter *et al.* 1998).

1.11. Objectives I - identifying novel vaccine candidates

The need to develop an effective sub-unit vaccine against tropical theileriosis has been highlighted. While current vaccine candidate genes show encouraging preliminary results, further studies are required to confirm their suitability for use in the field. Significant polymorphism at the amino acid level has been detected in the antigens presently characterised. However, it is currently unclear exactly what role antigenic diversity plays (a) in allowing the parasite to evade or modulate the host immune response and (b) in influencing the population structure of the parasite in field populations. The high level of polymorphism in *TaMS1* alleles has been related to the positive, diversifying pressure from the bovine immune system (Katzer *et al.* 1998; Gubbels *et al.* 2000b). It is reasonable to speculate that this may be a general feature of *T. annulata* antigens since extensive polymorphism has been encountered in antigens in other *Theileria* species, as described above, and also in related apicomplexans such as the *Plasmodia*. In order to facilitate the development of an effective sub-unit vaccine, further candidate genes require identification, particularly those stage-specifically expressed in the macroschizont. However, before a gene may be regarded as a suitable candidate for vaccination, it first must be identified as encoding an antigenic protein. Therefore, the identification of a

subset of highly diverse genes in the genome of *T. annulata* would be a logical starting point for such a quest. The availability of sequence data representing both *T. annulata* and *T. parva* genomes provided the perfect resource for such an investigation. A primary objective of this study was to use these resources to identify novel antigen genes and investigate their allelic diversity in *T. annulata*. Specifically the following aims were defined -

1. To determine whether comparative genomics can be used to support the hypothesis that *Theileria* surface antigens are generally under the influence of diversifying selection.

2. To bioinformatically screen the entire genome of *T. annulata* to identify a panel of merozoite and / or macroschizont genes that may encode protective antigens.

3. To determine if these putative antigens exhibit polymorphism at the species level, by sequencing a representative number of alleles and to assess whether this polymorphism is consistent with immune selection.

4. To compare and contrast sequence diversity of novel predicted antigens with the current merozoite vaccine candidate TaMS1.

5. To identify a limited number genes, which may be considered as vaccine candidate antigens.

To summarise, the rationale of this approach is to use the relatively rapid and inexpensive method of bioinformatically mining the genome before using statistical and molecular techniques to predict a limited number of suitable genes. This represents an effective use of time and resources, whereby a few promising vaccine candidates can be fed forward into a laboratory and field based vaccine development programme. It is anticipated that in turn, immunochemical and ultimately field trials will support the validity of this approach.

1.12. Population genetics

1.12.1. Introduction

Studies on antigen genes have revealed that there is considerable genetic diversity among isolates of *T. annulata*. In general, these studies have been directed toward analysing the nature of polymorphism at a single locus and relating this to the structure and function of the encoded product. Often this is of direct practical significance when considering genes

as sub-unit vaccine candidates. However, using *ad hoc* collections of isolates to analyse genes that may also be under selective pressure, provides a limited amount of information on genome-wide variation in natural populations of the parasite. To properly investigate this, suitably structured collections of isolates need to be genotyped using appropriate molecular markers. However, only a limited number of studies have specifically addressed the population genetics of *Theileria* species.

1.12.2. Studies in *T. annulata*

An extensive study was conducted in Tunisia (Ben Miled *et al.* 1994) to gauge the level of diversity of *T. annulata* stocks within a limited geographical area. The degree of polymorphism in defined geographical regions was assessed to determine whether diversity was localised over larger regions or whether high levels of variation could exist within a restricted area. 51 *T. annulata* stocks were collected from 17 different sites, within four bioclimatic zones, and comprised mainly macroschizont-infected cell lines established *in vitro*. The remainder of the stocks were piroplasm preparations purified from infected erythrocytes collected from infected animals. The isolates were examined for their Glucose Phosphate Isomerase (GPI) phenotype. Isoenzymes are variants of the same enzyme, encoded by different alleles at the same locus and because of amino acid charge variation, allozymes can be differentiated by their relative migration during starch gel electrophoresis. Many enzymes are invariant within populations and most polymorphic enzymes have only a few variants. Although this limits the power of isoenzyme analysis to resolve genetic differences, these markers are time and cost efficient for research requiring only a few polymorphic proteins. The majority of isolates in Ben Miled's study showed a multiple triplet band pattern with seven variant forms identified.

In addition, two further analysis were performed on the isolates: (a) restriction fragment length polymorphism was detected by Southern blotting genomic DNA with the single-copy gene probes *Ta*T 17 (5 genotypes) and *Ta*T 21 (3 genotypes) and (b) antigenic polymorphism was detected by IFAT using mAb 7E7, with the percentage of macroschizonts reacting to this antibody being recorded.

Study of the infected cell lines demonstrated that the frequency of different parasite genotypes and phenotypes was variable and mixtures were more frequent than homogeneous populations. This suggestion of mixed parasite populations was consistent with previous studies (Shiels *et al.* 1986; Williamson *et al.* 1989). In one of these studies, variation between *T. annulata* isolates from different countries was shown using

monoclonal antibodies (Shiels *et al.* 1986). In the Ben Miled study, a very high level of polymorphism was demonstrated within Tunisia with no characteristics common to stocks isolated from throughout the country or from within each bioclimatic zone. This finding was later echoed in the TaMS1 diversity study (Gubbels *et al.* 2000b) where no evidence of geographical sub-structuring could be detected.

In Tunisia a similar level of polymorphism was found within sites and farms of each zone (Ben Miled 1993; Ben Miled *et al.* 1994). These results suggest that the Tunisian population of *T. annulata* has a single panmictic structure, without any obvious level of sub-structuring, although this hypothesis was not formally tested. Furthermore, the designation of farms as being within defined bioclimatic zones may have been a drawback in this study, as it may have been more informative to group farms according to physical proximity, even though it is likely the results would have been identical. The bioclimatic zoning approach is probably more useful when examining disease epidemiology as it relates to the state of disease endemicity and local vector ecology (see Section 1.7.).

1.12.3. Studies in *T. parva*

The molecular epidemiology of *T. parva* has been investigated using a combination of RFLP techniques (Geysen *et al.* 1999). Southern blots on RFLP-DNA using '*Tpr*' (*Theileria parva* repeat) and telomere probes determined relative homogeneity among Zambian isolates, contrasting with high heterogeneity in previous Kenyan studies (Conrad *et al.* 1987; Bishop *et al.* 1997). The Geysen study did however agree with results from cattle in Zimbabwe (Bishop *et al.* 1994b) and RFLP-PCR assays for three loci (*PIM*, *p104* and *p150*) revealed PIM was most polymorphic, with ten alleles identified. Using RFLP-PCR at the *PIM* locus, isolates from two geographical areas in Zambia could be differentiated in addition to provinces within that country. However, both *p104* and *p150* were monomorphic for the same isolates, notwithstanding a single outlier. Taken together, the data suggest a relatively homogeneous population in Zambia, which can only be discriminated by a single highly polymorphic antigen gene. Moreover, the field data strongly suggested that an introduced vaccine stock had widely disseminated in one province, following live immunisation. The results indicate a homogeneous, epidemic structure whereby the clonal expansion of an introduced genotype has come to dominate the population. The authors explain the results by suggesting that prior to immunisation the population consisted of a high number of naïve hosts, which together with optimal tick conditions laid the ground for an epidemic.

The completion of the genome sequence of *T. parva* has allowed the development of a new generation of genetic markers for population studies. Using this resource, a panel of 11 micro-satellite and 49 mini-satellite polymorphic markers was identified (Oura *et al.* 2003). These fast evolving markers were distributed across all four chromosomes and were shown to be specific to *T. parva*. In a preliminary study, up to eight alleles were detected per locus and a high level of diversity was displayed within a single Kenyan population. There was no clear evidence of a correlation between geographical origin and relatedness of parasite stocks.

To investigate the population structure in three geographically distinct areas, Oura *et al.* applied a subset of this panel of markers to field populations in Uganda (Oura *et al.* 2005). Three districts where ECF was endemic were selected – Lira, Mbarara and Kayunga – each separated by around 300 km. Blood samples were collected from adult cattle and calves and immobilised on FTA filters. PCR amplification of twelve micro- and mini-satellite loci was subsequently carried out on these blood spots using pairs of nested primers and the resulting products analysed on high-resolution 'Spreadex' gels. For each locus in every sample, the predominant allele was identified on the basis of ethidium bromide staining intensity. This was made easier since a polymorphic mini-satellite had already been used to screen all blood samples, biasing the dataset towards younger animals with a less heterogeneous parasite burden. 81% of adult cattle showed multiple genotypes with this marker, compared with 35 % of calves, which is likely to be related to limited exposure to challenge in the younger stock. Multi-locus genotypes (MLGs) for each sample were then generated using the predominant allele, identifying 84 unique MLGs were from a total of 104 samples. A high level of diversity was found in each region, with the samples from Kayunga more similar to themselves than to isolates from the other two areas. Limited genetic differentiation was demonstrated between Mbarara and Lira, although a sub-group consisting exclusively of Mbarara isolates was clearly identified. A larger less well-defined Lira sub-group was also identified. Linkage disequilibrium (LD) was evident when all the samples were treated as one population implying a degree of sub-structuring in the population, which may be ascribed to geographical and/or genetic isolation. Interestingly, LD was still demonstrated when each of the three populations were analysed in turn. However, when identical isolates from Lira were treated as one, LD disappeared suggesting an epidemic structure in this group. Linkage equilibrium (LE) was also demonstrated in the Mbarara group when both the sub-group and the set of identical samples were removed from the analysis. However, the Mbarara population sample was considered a genuine cohort group, where the animals were managed identically as a unit.

There was therefore no basis for excluding them from the analysis *a priori* thus suggesting an unidentified mechanism for sub-structuring within this group. Unexplained sub-structuring was also identified in Kayunga, however the sample size was smaller. A spectrum of population structures has been documented in *P. falciparum* where epidemic populations have been correlated with areas of reduced disease transmission (Anderson *et al.* 2000a). Differences in transmission rates of ECF may therefore explain some of the slightly obscure findings in this study. A lack of clear geographical sub-structuring between populations, may be related to local cattle movement in this study, moreover the finding is not entirely unsurprising given that other studies in both *T. annulata* and *T. parva* have been unable to detect such a relationship.

1.12.4. Genetic exchange and recombination in the vector

A significant conclusion from the Oura study is that genetic exchange occurs frequently in *T. parva* (Oura *et al.* 2005). This is further evidence supporting the existence of a sexual stage in the parasite life-cycle, since a high level of recombination between loci indicates frequent crossing over during meiosis. Intragenic recombination in *TaMS1* (Gubbels *et al.* 2000b) supports the theory that sexual recombination is taking place within *T. annulata* populations while frequent sexual recombination between widely distributed genotypes would explain the inability to geographically cluster isolates based on *TaMS1* sequence diversity. While the data on *T. annulata* are consistent with genetic recombination occurring at high frequency and a lack of geographical sub-structuring, this question has not been directly addressed and recombination has not been formally demonstrated. Based on the morphological data, previously discussed in Section 1.3.2., meiosis in *T. annulata* is presumed to take place in the tick gut (Schein and Friedhoff 1978). In order to demonstrate sexual reproduction, Ben Miled *et al.* co-infected ticks with different cloned strains (unpublished data). Recombinant genotypes were identified demonstrating the potential for genetic exchange and recombination between different parasite sub-populations in the field. In addition, sexual recombination between stocks of *T. parva* has been demonstrated in the laboratory (Bishop *et al.* 2002).

From experimental infection of calves with populations of *T. annulata* macroschizont infected cell lines, it was noticed that changes occurred in both the relative proportions and nature of genotypes recovered from infected animals (Ben Miled 1993). A series of five calves were infected with different clonal genotypes and four additional calves were infected using a pool of all five clones. Only one of the five parasite clones failed to establish in the calf infected with that clone alone. The same clone also failed to establish

in the four pooled infections. This suggests there is a degree of selection taking place, perhaps with some variants being better adapted to the bovine host. Alternatively, it may be taken as evidence that certain clones had adapted to culture conditions. Nevertheless, it did suggest that there is variability between parasite genotypes in their ability to establish in cattle. This in turn raised the possibility that the population of *T. annulata* within a calf could be different from the reservoir from which it was infected, i.e. the macroschizont-infected cell line in this experiment and possibly the tick vector in the field. In addition, it appeared that mixed infections lead to more severe clinical disease than clonal infections. The cattle infected with the pool of clones exhibited a higher parasitaemia than calves infected with a single clone. Although other clinical parameters were very similar, the study suggested a degree of synergy between individual genotypes. Furthermore, when nymphs were fed on the infected calves, the infection rate was dramatically higher in the ticks fed on calves subjected to pooled infection, with a much higher mean number of acini infected. Using several genetic markers, parasite diversity was analysed in (a) cell lines derived from PBMs of infected calves, (b) cell lines derived ticks fed on these animals and (c) tick salivary gland lysates. To summarise, each clone did not behave identically and a spectrum of efficiency of transmission was encountered. However, it must be stressed that this experiment was based on a very limited number of observations, but does provide preliminary evidence for variation in phenotypes such as transmission, infectivity and virulence.

The population structure of *T. annulata* was investigated during bovine infection and following transmission to ticks (Gubbels *et al.* 2001). During persistent infection derived from sporozoite inoculation of four calves, *TaMS1* allelic diversity was assessed using PCR product sequence polymorphism visualised by denaturing gel electrophoresis. In the experiment, ticks were fed on these animals soon after infection (acute stage) then two months later (carrier stage). Three *TaMS1* alleles were initially identified in the bovine blood. The infection profile of the ticks fed on acutely infected calves was very similar to the profile generated from the blood itself. However, in the case of a single calf, the tick samples obtained from the carrier state only displayed single genotypes, which were distinct from the blood profile at this point. Several different single genotypes were detected in these ticks, including a further novel *TaMS1* allele identified in one tick. The study demonstrates there is little alteration in detectable genotypes through the course of the bovine infection, but carrier animals may be limited in the number of genotypes they can transmit. The authors suggest selection may be the result of specific inhibition of tick transmission of carrier-phase dominant parasites, mediated by a humoral or cellular

response. Inhibition would occur once infected erythrocytes are lysed within the tick gut and thus the piroplasms/gametes exposed to such immune mechanisms.

The analysis undertaken by both Ben Miled and Gubbels *et al.*, while of a necessarily preliminary nature, suggests that the population dynamics of mixed genotype infections in calves is complex, that there are interactions between different genotypes and that there are biological differences between the different parasite genotypes.

1.12.5. Summary

In light of studies on *T. annulata* to date, it appears there are two major gaps in knowledge regarding the basic population biology of the parasite -

1. Sexual recombination, which has been predicted to occur in the tick gut, has not been formally demonstrated either experimentally or with neutral genetic markers in natural populations.

2. The role of genetic exchange to generate diversity in field populations of the parasite has not been quantified. This, and the possibility of sub-structuring of geographically isolated populations has not been refuted or confirmed using multiple polymorphic markers. Is the underlying population structure clonal, panmictic or an intermediate epidemic structure where sexual recombination is present, but concealed by clonal expansion of a limited number of genotypes?

1.13. Objectives II - determining the underlying population structure of *T. annulata*

Although the work of Ben Miled (Ben Miled 1993; Ben Miled *et al.* 1994) is suggestive of a single, frequently mating population of *T. annulata* within Tunisia, a study focussing on the population structure of the parasite in the field employing more discriminatory markers is required. A panel of micro- and mini-satellite markers, analogous to those developed for genotyping *T. parva* (Oura *et al.* 2003) would be ideal for such a purpose. Furthermore, it would be useful for such a study to encompass an entirely separate geographical area to test if conclusions about population structure in one country can be applied to other countries affected by the disease. With this in mind, the following objectives were defined –

- Development of a panel of neutral genetic markers to create a multilocus genotyping system specific for *T. annulata*.

- Determination of the population structure of *T. annulata* in two geographically distinct areas. The North African country of Tunisia was selected since it would allow comparisons to be drawn in relation to previous work. The second focus of study is the Western coastal region of Turkey around Izmir. This region has similar epidemiology to that of Tunisia, however a distance of around 1,500 kilometres separates the two areas. With the geographical barriers of the Mediterranean Sea, mountain ranges and the various North African and Middle Eastern countries dividing them, limited transit of livestock and vector would be anticipated.

Specifically, the work presented in this thesis set out to ask the following questions –

(a) What is the underlying population structure, with reference to sexual recombination?

(b) Is there geographical sub-structuring of the parasite population between and within countries?

(c) How does host phenotype (i.e. breed, sex and age) relate to parasite diversity?

(d) Can differences in parasite populations be detected between areas using different approaches to theileriosis control?

CHAPTER TWO

DEVELOPMENT AND APPLICATION OF NEUTRAL MARKERS

2.1. Introduction

2.1.1. Population structure

A range of population structures has been demonstrated in studies of apicomplexan species, ranging from clonal in *Toxoplasma gondii* (Boyle *et al.* 2006) to panmixia (random mating) in a major zoonotic subtype of *Cryptosporidium parvum* (Mallon *et al.* 2003a). Population structure not only varies between different species but also within a single species in different ecological situations. For example, it was demonstrated that there are differences in the population structure of *Plasmodium falciparum* in different geographical locations with significant linkage disequilibrium (LD) observed in six of nine populations studied, corresponding to regions where the transmission rate was low (Anderson *et al.* 2000a). Identical multilocus genotypes (MLGs) were present in these populations, suggesting a high level of self-fertilisation, while populations from regions with high transmission rates were shown to be panmictic. In contrast to this trend, other studies have indicated that regions with a high rate of transmission are not panmictic and show evidence of inbreeding with linkage disequilibrium (Razakandrainibe *et al.* 2005) and inbreeding in the absence of detectable linkage disequilibrium (Paul *et al.* 1995). Although it is accepted that *P. falciparum* has an obligate sexual stage (Talman *et al.* 2004; Baton and Ranford-Cartwright 2005; Cowman and Crabb 2005), some researchers argue that the underlying population structure of the parasite remains clonal (Rich *et al.* 1997; Urdaneta *et al.* 2001; Razakandrainibe *et al.* 2005). However, the balance of evidence suggests that sexual recombination plays an important role in generating genotypic diversity and in general, levels of inbreeding inversely correlate with local transmission intensity.

The population genetics of *T. parva* have recently been investigated using a panel of polymorphic markers to analyse isolates from different regions of Uganda. A high level of diversity was described, with many isolates consisting of a mixture of genotypes and linkage disequilibrium was initially identified in three populations isolated from different areas. However when isolates with identical genotypes were treated as a single isolate, two

of the populations were found to have an epidemic structure, where genetic exchange occurs frequently but is masked by over representation of some MLGs (Oura *et al.* 2005). This study is covered in some detail in Section 1.12.3.

To date, a detailed population genetic analysis of *T. annulata* has not been undertaken and thus little is known about the role of genetic exchange in this parasite. However two sets of studies have demonstrated significant levels of polymorphism both within and between isolates and indicated that sexual recombination may be occurring. In the first set of studies, diversity of *T. annulata* stocks was analysed using isoenzymes, RFLPs of two single copy gene probes (Ben Miled *et al.* 1994) and monoclonal antibodies (Shiels *et al.* 1986). Analysis of 51 stocks of *T. annulata* from Tunisia (Ben Miled *et al.* 1994) assessed the degree of polymorphism in four bioclimatic zones (17 different sites) and examined whether diversity was localised to some of these regions and whether high levels of variation existed within a restricted area. The frequency of different parasite phenotypes/genotypes was found to be variable across the regions studied and mixtures of different parasite strains were more frequent than homogeneous populations, agreeing with previous studies where mixed parasite populations were identified (Shiels *et al.* 1986; Williamson *et al.* 1989). Given the lack of obvious geographical sub-structuring, the results suggested that within Tunisia, *T. annulata* exists as a single population. In the second set of studies, alleles of two surface antigen genes were sequenced from a number of isolates and significant levels of polymorphism identified. The most extensively studied antigen gene, the merozoite and piroplasm antigen (*TaMS1*) displays significant sequence diversity within and between isolates (Gubbels *et al.* 2000b). Phylogenetic analysis of the sequence polymorphism showed no geographical clustering of sequence variants and almost identical sequences occurred in different geographical regions, leading the authors to suggest that the population structure is panmictic, although this was not formally tested. The single copy sporozoite surface antigen gene (*SPAG*), has also been shown to exhibit a high degree of polymorphism by comparing two full-length alleles (Katzer *et al.* 1994) with RFLP analysis of further isolates identifying further allelic variants. Together these studies demonstrate a high level of diversity between different parasite isolates, however the question remains - is a high level of sexual recombination the primary mechanism for the generation of genotypic diversity within these parasite populations? To answer this question, a novel set of suitable genetic markers needs to be identified in order to undertake formal population genetic analysis.

2.1.2. Micro- and mini-satellite genotyping

Micro- and mini-satellite genotyping provides a convenient method for genetic studies of *T. annulata*. Such markers have already been successfully applied to determining the population structure of several species of apicomplexan parasites (Anderson *et al.* 2000a; Mallon *et al.* 2003a; Oura *et al.* 2005). In the *Theileria* genus, mini-satellite markers were first developed in *T. parva* (Bishop *et al.* 1998). These putatively neutral loci are composed of tandem repeats of short DNA motifs (7 - 24 bp), which generally are subject to a very high frequency of mutation in motif copy number (Jeffreys *et al.* 1988). This has facilitated the development of multilocus genotyping systems capable of detecting variation between individuals. Such systems have been used extensively in population genetic applications and in human forensic medicine (Gill *et al.* 1985; Jeffreys *et al.* 1985a). Screening a *T. parva* λgt11 genomic library with the mini-satellite regions of bacteriophage M13 lead to the identification of several widespread mini-satellite loci (Bishop *et al.* 1998). Radiolabelled probes detected more than 20 loci, two of which were analysed in isolation and found to be able to differentiate between two different *T. parva* isolates. One of these single mini-satellite probes was used to analyse Zambian and Kenyan stocks of *T. parva* in conjunction with a variety of other molecular markers (Geysen *et al.* 1999). This was a relatively crude RFLP method and provided only a fingerprint representation of each isolate. Multiple bands indicated the presence of the motif in a variety of genomic locations after restriction enzyme digestion. Genetically, the data are almost impossible to interpret, as the appearance of each band relies on the presence of a particular repeat region and conservation of specific restriction sites and consequently the relatedness of individual fingerprints to one another cannot be determined. Identifying specific mini-satellite loci and measuring their individual polymorphism is necessary to capture the genetic information they contain. However, it was not until the advent of the *T. parva* genome sequence that these loci could be catalogued and screened *en masse*. By bioinformatically scanning this database, a panel of 49 polymorphic mini-satellites and 11 polymorphic micro-satellites was identified (Oura *et al.* 2003). Like mini-satellites, micro-satellites are neutral polymorphic molecular markers, however they have a shorter period length of between two and six bases. Micro-satellites owe their variability to a high rate of mutation by slipped strand mis-pairing (slippage) during DNA replication. The rate of such mutations is significantly higher than the rate of base substitutions and may vary between 10^{-6} and 10^{-2} per generation (Schlotterer 2000). Mutation may also occur during meiotic crossing over (Blouin *et al.* 1996), although this mechanism is more usually associated with the even higher level of diversity observed in

mini-satellites. The mechanism of DNA slippage can occasionally lead to incorrect amplification of micro-satellites if it occurs early on during a PCR reaction, resulting in fragments of incorrect size. Similar to mini-satellites, micro-satellites are considered to be abundant and widespread over eukaryotic genomes (Tautz and Renz 1984), a feature certainly displayed by the *T. parva* markers. In the study by Oura *et al.* in 2003, primers were designed to the flanking sequence to permit PCR amplification of the markers after which products were separated using electrophoresis on 2 % agarose gels. For some markers this did not provide enough resolution to differentiate alleles of similar size and therefore high-resolution 'Spreadex' gels, which can discriminate differences as little as 3 bp were utilised. Using this system, these markers were capable of documenting extensive diversity in a limited number of *T. parva* isolates and subsequently they were applied to field population studies in Uganda (Oura *et al.* 2004; Oura *et al.* 2005) and Kenya (Odongo *et al.* 2006). The successful application of these markers in a closely related species coupled with the recent publication of the genome sequence of *T. annulata*, suggests a similar panel of polymorphic markers could be identified and developed in this species. Furthermore, the recent availability of capillary-based sequencers provides an improved method for separating and identifying PCR products at very high resolution. The use of such a system is imperative to extract the maximum amount of information from heterogeneous field isolates, containing a multiplicity of potentially closely related genotypes.

2.1.3. Multiplicity of infection

It is anticipated that a large amount of heterogeneity will be present in field isolates of *T. annulata*. This presents a challenge, since DNA preparations of the blood of infected animals will contain multiple alleles at each locus. How, then can one distinguish between different haploid genotypes of the parasite? The majority of the DNA available for this study was from cell lines derived from the blood of infected cattle. These low passage cultures (p9 or p10) may also be expected to contain a mixture of genotypes. The effect of prolonged maintenance *in vitro* is known to attenuate the parasite, both in terms of virulence and in its ability to differentiate into merozoites (Hall *et al.* 1999). Previous studies have suggested this attenuated phenotype may be as a result of a reduction in the number of genotypes (Darghouth *et al.* 1996a) and / or alteration in parasite gene expression (Sutherland *et al.* 1996; Hall *et al.* 1999). To test whether there was a correlation between high growth rate *in vitro* and the loss of ability to differentiate to the merozoite stage, a correlate of attenuation, a recent study compared the growth rate of clones at 37 °C with the rate of their differentiation at 41 °C (Taylor *et al.* 2003). A panel

of 22 genetically distinct clones was selected on the basis of PCR-RFLP profile and mAb reactivity and with one exception, differentiation correlated positively with growth rate. These observations do not therefore explain the attenuation and reduction in diversity displayed by prolonged culture of *T. annulata*. However, the general trend towards a reduction in heterogeneity was demonstrated in *T. annulata* by comparing genetic diversity in piroplasm extracts with derived cell lines (Ben Miled *et al.* 1994). Five paired sets of DNA were analysed using three markers: GPI iso-enzyme phenotype and two polymorphic loci visualised by Southern blotting. Three of the sets showed a higher level of diversity in the piroplasm extract, one was equal and one showed higher diversity in the cell line. For the latter, four GPI types were identified in the cell line, while only one of these was found in the piroplasm. For some homologous sets, a novel probe profile or GPI phenotype was identified in the cell line, suggesting the amplification of a minor genotype. To complement this study, it would be useful to genotype matching pairs of cell line and piroplasm DNA preparations with a newly developed multilocus genotyping system. This technique relies on identifying the allele present at each locus and compiling the data across several loci. This is most easily applied to parasite systems where a single strain can be isolated from the host, such as in the case of *C. parvum* (Mallon *et al.* 2003a). For organisms such as *T. parva* and *P. falciparum*, the most convenient method of obtaining parasite genetic material from the field is by sampling blood from the infected host population. Because these samples represent a mixture of haploid genotypes, multiple alleles are detected at each locus when genotyping with polymorphic micro- and mini-satellite markers (Anderson *et al.* 2000a; Oura *et al.* 2005). In order to undertake population genetic analysis, a discrete haploid genotype must be ascertained for each sample to permit standard measurements of population differentiation, linkage disequilibrium and the calculation of allele frequencies. As in previous studies, this is achieved by determining the most abundant allele at each locus and combining the data to form a MLG, which represents the most abundant haploid genotype in the mixture (Anderson *et al.* 2000a; Oura *et al.* 2005).

2.1.4. Aims of this chapter

The objectives of the work presented in this chapter were to identify, characterise and use a panel of polymorphic micro- and mini-satellite markers to analyse diversity in field populations of *T. annulata*. This panel of markers was then used to address a series of fundamental questions about the population genetics of this parasite, namely -

1. **What is the population structure of Turkish and Tunisian stocks of *T. annulata*?** Using standard population genetic techniques, the hypothesis that *T. annulata* has a panmictic population structure is formally tested.

2. **Can genotypes from widely separate geographical locations be differentiated, or do all these isolates comprise a single population?** Samples from Tunisia, Turkey and small number of stocks from Southern Europe, North Africa and the Middle East are analysed to test if geographical sub-structuring can be detected.

3. **Can selection be detected by comparing the multiplicity of infection in cattle with *in vitro* heterogeneity?** This is achieved by genotyping piroplasm extracts and homologous cell lines cultures and comparing the number of alleles identified at each locus. This is supported by detailed analysis of cell lines and clones derived from the *T. annulata* (Ankara) isolate.

These questions are critical for predicting and analysing the response of *T. annulata* to selective pressures such as vaccination and drug treatment both locally and in the wider context.

2.2. Materials and methods

2.2.1. Parasite material and DNA preparation

The collection of *Theileria* parasite stocks used in this study represents Tunisian and Turkish isolates in addition to eight stocks derived from the other countries, as detailed in Table 2.1. The distribution of sampling sites in Tunisia and Turkey is presented in Figure 2.1. Most of the Tunisian stocks were low passage macroschizont infected cell lines (p8 - p10), which were established from peripheral blood mononuclear cells (PBMs) isolated from cattle suffering clinical disease (Ben Miled *et al.* 1994). Clones were produced by the limiting dilution of these cell lines using a previously described method (Shiels *et al.* 1992). Cryopreserved macroschizont infected cell line stocks were thawed and cultured in 25 cm^2 tissue culture flasks using Roswell Park Memorial Institute (RPMI) 1640 medium, supplemented with 15 % foetal calf serum. Approximately 10^7 infected cells were centrifuged at 1500 g for five minutes and the cell pellet washed in phosphate buffered saline (PBS). The final cell pellet was resuspended in PBS from which DNA was purified using a Qiagen QIAamp DNA Mini Kit according to the manufacturers instructions. The Tunisian piroplasm samples were obtained by lysing erythrocytes from

Table 2.1. *T. annulata* stocks used in population genetic analysis

51 parasite stocks were analysed from Tunisia consisting primarily of cell lines maintained *in vitro*. They were isolated from infected animals in the North-eastern, North-western and Central regions of the country (see Figure 2.1.(i)). Full details of these isolates can be found in the PhD thesis of Leila Ben Miled (Ben Miled 1993). 13 Turkish stocks were analysed from Ankara province in Anatolia (3 cell lines, 3 clones) and Aydın province in the Aegean region (7 cell lines). The location of Aydın and Ankara are detailed in Figure 2.1.(ii). Eight cell lines from other countries were also analysed, but were not included in much of the formal population genetic analysis.

Table 2.1. *T. annulata* stocks used in population genetic analysis

	Stage	Type	Area of isolation	Reference
Tunisia (51)				
3 isolates	schizont	cell line	Humid zone (North-western Tunisia)	Ben Miled 1993
2 isolates	schizont	cell line	Sub-humid area (Northern Tunisia)	"
40 isolates	schizont / piroplasm	cell line / field isolate	Semi-arid area (North & Central Tunisia)	"
6 isolates	schizont	cell line	Arid area (Central Tunisia)	"
Central Turkey (6)				
T.a. Ankara A46 clone 4	schizont	clone *	Ankara, Ankara province	*(See Section 2.2.1)*
T.a. Ankara A2 D3	schizont	clone	Ankara, Ankara province	*(See Section 2.2.1)*
T.a. Ankara A2 D7	schizont	clone	Ankara, Ankara province	*(See Section 2.2.1)*
T.a. Aldere calf 53	schizont	cell line	Aldere, Ankara province	-
T.a. Pendik †	schizont	cell line	Ankara, Ankara province	Özkoc and Pipano 1981
T.a. Abidinpaşa	schizont	cell line	Abidinpaşa, Ankara province	-
South-western Turkey (7)				
T.a. Akçaova	schizont	cell line	Akçaova, Aydın province	Ilhan 1999
T.a. Dalama-Kozak	schizont	cell line	Dalama, Aydın province	"
T.a. Cine	schizont	cell line	Cine, Aydın province	"
T.a. Haci Ali Obasi	schizont	cell line	Incirli, Aydın province	"
T.a. Aydın	schizont	cell line	Germeçik, Aydın province	"
T.a. Koçarli	schizont	cell line	Koçarli, Aydın province	"
T.a. Yenihisar	schizont	cell line	Soke, Aydın province	"
Other countries (8)				
T.a. Shambat 32	schizont	cell line	Sudan	Shiels *et al.* 1986
T.a. Shambat 33	schizont	cell line	Sudan	"
T.a. Soba (2A5)	schizont	cell line	Sudan	"
T.a. Sagadi	schizont	cell line	Sudan	"
T.a. Razi S3	schizont	cell line	Iran	Hooshmand-Rad and Hashemi-Fesharki 1968
T.a. Tova	schizont	cell line	Israel	Pipano 1974
T.a. Caceres	schizont	cell line	Spain	De Kok *et al.* 1993
T.a. Ode	schizont	cell line	India	Baylis *et al.* 1992

† vaccine line developed at Pendik institute, but isolated from Ankara province

* represents cloned macroschizont infected cell line

Figure 2.1. Sampling sites used in the preliminary study

(i) Tunisia

Parasite material was collected by Leila Ben Miled from 18 farms in northern and central Tunisia. Sampling sites are indicated by small black circles; a cluster of eight farms is identified by the larger circle (adapted from Ben Miled, 1993).

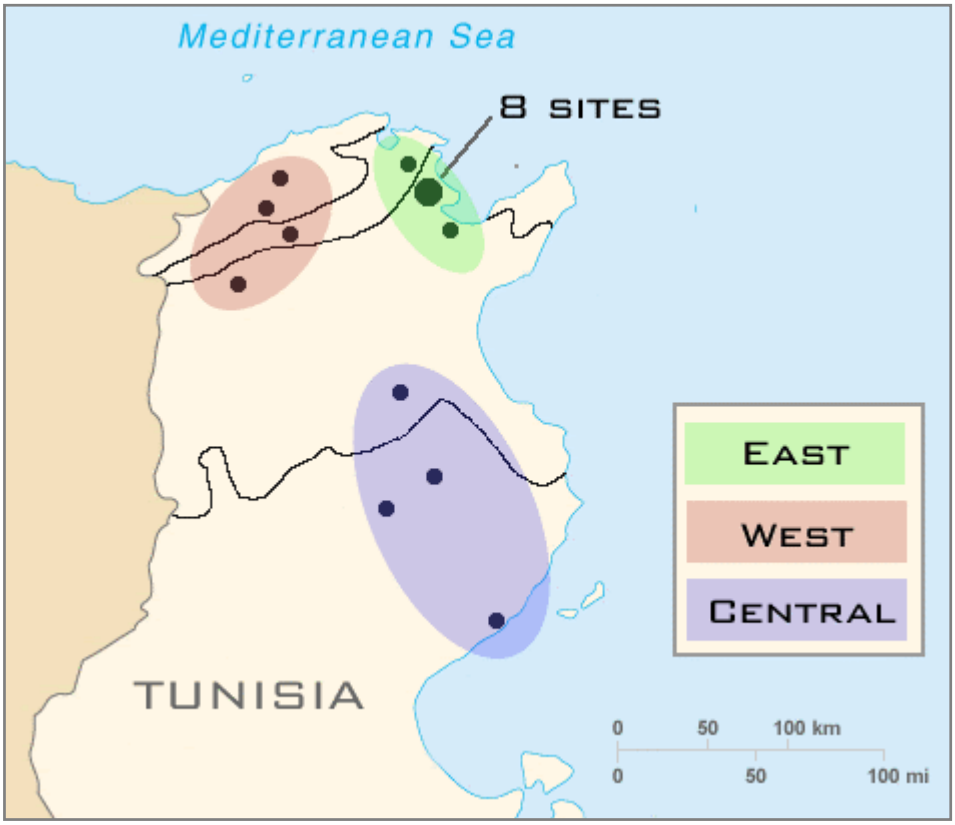
(ii) Turkey

Turkish samples were collected from the province of Ankara in central Anatolia and from the province of Aydın in the Aegean coastal region. These represent two geographically distinct areas separated by a distance of approximately 500 km.

Details of the samples collected in each country are presented in Table 2.1.

Figure 2.1. Sampling sites used in the preliminary study

(i) Tunisia



(ii) Turkey

an infected animal with ammonium chloride and subsequently purifying the piroplasms by centrifugation (Conrad *et al.* 1987).

For the initial screening of markers, DNA from a panel of 18 stocks of *T. annulata* was used and included six Tunisian stocks representing isolates from the four bioclimatic zones (Ben Miled *et al.* 1994), five Turkish stocks isolated from Central (Ankara, Sarioba, Abidinpaşa), Western (Balıkesir) and North-western Turkey (Bursa) areas as well as DNA from three piroplasm preparations – Ankara (Turkey), Hissar (India) and Maroc (Morocco). These are documented in Table 2.2. Samples from diverse geographical origins were chosen to maximise the probability of detecting polymorphism, while samples within and between areas in Tunisia were included to evaluate the level of polymorphism within a putatively sympatric population. A strain of *T. sergenti* (Dr S. Kawazu, NIAH Japan) was also incorporated in this panel to test for the species specificity of the markers. Three Tunisian cloned infected cell lines were also included to determine whether a single locus was amplified with each primer pair. As *T. annulata* is haploid, a single allele would be predicted if the sequence was single-copy. DNA from several other species of *Theileria* were also included in the study – *T. parva* (two preparations – Muguga and Marikebuni (Morzaria *et al.* 1995)), *Theileria* sp. (China) (Schnittger *et al.* 2000), *T. taurotragi* (CVTM, Edinburgh) and *T. lestoquardi* (Hooshmand-Rad 1985) to determine the species specificity of the markers.

The *T. annulata* (Ankara) isolate and a series of its derivatives were analysed in order to investigate *in vitro* selection. Following isolation near Ankara in Turkey, a cell line was transported to Berlin and used to experimentally infect *Hyalomma anatolicum anatolicum* ticks (Schein *et al.* 1975). A batch of infected ticks was transported to the Centre for Tropical Veterinary Medicine (CTVM) in Edinburgh where the Ankara stock was maintained by alternate passage in ticks and cattle. A piroplasm preparation was made from blood from an infected animal, using the method previously described. The *T. annulata* A_2 cell line was derived by *in vitro* infection of peripheral blood monocytes (PBMs) from 'calf 2' using a *T. annulata* (Ankara) sporozoite stabilate. Seven clones derived from the *T. annulata* A_2 cell line were generated by the limited dilution technique (Shiels *et al.* 1992) and were designated A_2 C9, A_2 D7 etc.

2.2.2. Identification of tandemly repeated sequences

In August 2002, the *T. annulata* genomic sequence was available as a number of large contigs. In order to identify micro- and mini-satellite loci, these data were downloaded

Table 2.2. Panel of stocks and isolates used to screen loci

For the initial screening of markers, DNA from a panel of 18 stocks and isolates of *T. annulata* was used as template for PCR amplification. This panel was compiled to represent clonal, cell line and piroplasm parasite samples and was designed to test (a) whether each marker represented a single locus and (b) if mixed infections could be detected. Samples from diverse geographical origins were chosen to maximise the probability of detecting polymorphism. Additionally, DNA representing an isolate of *T. sergenti* was included to test for species specificity.

Table 2.2. Panel of stocks and isolates used to screen loci

	Stock / isolate	Stage	Type	Country	Reference
1	*T.a.* Ankara	piroplasm	field isolate	Turkey	*(See Section 2.2.1.)*
2	*T.a.* Maroc	piroplasm	field isolate	Morocco	Ouhelli 1985
3	*T.a.* Hissar	piroplasm	field isolate	India	Gill *et al.* 1976
4	*T.a.* A_2 37	schizont	cell line	Turkey	*(See Section 2.2.1.)*
5	*T.a.* A_2 D7	schizont	clone (from A_2 37 cell line)	Turkey	*(See Section 2.2.1.)*
6	*T.a.* Aldere calf 53	schizont	cell line	Turkey	-
7	*T.a.* Sarioba calf 89	schizont	cell line	Turkey	-
8	*T.a.* Abidinpasa	schizont	cell line	Turkey	-
9	*T.a.* Mustafa Kemal	schizont	cell line	Turkey	-
10	*Theileria sergenti*	schizont	cell line	Japan	*(See Section 2.2.1.)*
11	*T.a.* BAT cl4 clone 5	schizont	clone*	Tunisia	-
12	*T.a.* BV4 cl5 clone 2	schizont	clone	Tunisia	-
13	*T.a.* BV4 cl4 clone 3	schizont	clone	Tunisia	-
14	*T.a.* LBM-16	schizont	cell line	Tunisia (N. central)	Ben Miled 1993
15	*T.a.* LBM-30	schizont	cell line	Tunisia (N. central)	"
16	*T.a.* LBM-44	schizont	cell line	Tunisia (N. central)	"
17	*T.a.* LBM-46	schizont	cell line	Tunisia (Southern)	"
18	*T.a.* LBM-39	schizont	cell line	Tunisia (Coastal)	"
19	*T.a.* LBM-36	schizont	cell line	Tunisia (Coastal)	"

* subsequently shown to be a mixture of two parasite genotypes

from ftp://ftp.sanger.ac.uk/pub/pathogens/T_annulata/ and screened with the tandem repeat finder program (Benson 1999). Repeat motifs up to 500 bp were identified using stringent parameters for identifying matches, mismatches and indels in the sequence, corresponding to weight values of 2, 7 and 7 respectively for the local alignment algorithm.

2.2.3. PCR amplification of loci

Primers were designed to the unique sequence flanking each repeat and used to amplify DNA from the panel of stocks and isolates (Table 2.2.) to test for amplification and marker polymorphism (Table 2.3.). DNA preparations were PCR amplified in a total reaction volume of 20 µl under conditions described previously (MacLeod *et al.* 2000) using the following thermocycler conditions: 94 ºC for 2 minutes, 30 cycles of 94 ºC for 50 seconds, 50-60 ºC for 50 seconds and 65 ºC for 1 minute with a final extension period of 5 minutes at 65 ºC. The annealing temperatures of the primers for the ten markers selected for population analysis are detailed in Table 2.4. During the initial screen for polymorphism, the amplicons were separated by electrophoresis on 2 % agarose gels and stained with ethidium bromide. Gels were photographed under ultra-violet transillumination and the size of each PCR product determined with reference to either a 1 kb or 100 bp DNA ladder. An example gel image is presented in Figure 2.2., representing PCR products from the panel of 19 stocks and isolates amplified using primers TS15. High-resolution genotyping of the full set of stocks and isolates was undertaken by incorporating a fluorescently labelled (FAM) primer into the PCR reaction and separating the products of amplification with a capillary-based sequencer (ABI 3100 Genetic Analyser). DNA fragment size was determined relative to a set of ROX-labelled size standards (GS500 markers, ABI) using Genescan™ software, which allowed resolution of 1 base pair (bp) differences. An example electrophoretogram is presented in Figure 2.3., showing two PCR amplicons generated by amplification from a heterogeneous parasite DNA template using marker TS5. For all loci and DNA preparations, the fragment size (i.e. peak position) was determined to two decimal places. Analysis of the distribution of fragment sizes facilitated the creation of 'fixed bins' of variable size to score alleles. Multiple products from a single PCR reaction indicated a mixture of genotypes, allowing the generation of a total genetic 'fingerprint' representing the entire population of genotypes within each sample. The data from the sequencer output also provided a semi-quantitative measurement of the abundance of each allele, allowing the predominant allele to be identified and then used to generate a multilocus genotype (MLG) representing the most abundant genotype in each DNA preparation.

Table 2.3. Characterisation of markers

PCR primers were designed for 33 markers, which were then tested against a panel of 18 samples representing isolates of *T. annulata*, detailed in Table 2.2. The polymorphism for each marker was assessed both across the panel and within the Tunisian subset and the capability of each marker to detect mixed parasite genotypes was recorded (% mixed samples). Additionally, the predicted length of the PCR product for the C9 genome strain was calculated. The ten markers highlight in grey were selected for further analysis, since (a) they amplified a high proportion of DNA templates, (b) they were polymorphic and (c) they displayed a spectrum of PCR product sizes within the 200 – 500 bp range.

Table 2.3. Characterisation of markers

Marker	Total alleles	Tunisian alleles	% mixed samples	No. non-amplifying	Motif length (bp)	Repeat motif copy number	Genome - predicted length (bp)	Range (bp)
TS2	11	3	17	5	12	7.5	611	500-1000
TS3	4	0	22	11	12	8.1	263	200-300
TS4	9	4	39	2	10	9.7	261	200-300
TS5	5	4	33	1	6	13.8	282	200-300
TS6	7	6	6	0	10	13.6	389	400-500
TS7	6	3	22	4	27	11.6	455	100-200 & 400-500
TS8	9	8	39	0	12	11.2	306	200-350
TS9	8	6	39	0	3	27.3	366	300-400
TS10	4	2	6	9	11	12.5	458	200-500
TS11	8	2	6	9	41	7.6	411	200-450
TS12	9	4	33	1	6	10	267	200-400
TS13	7	4	22	4	11	11.8	228	<300
TS14	11	4	28	3	10	26.4	376	300-500
TS15	8	6	67	0	24	6.7	286	200-400
TS16	10	7	44	0	3	35	354	300-400
TS19	1	0	0	17	2	26.5	136	150 *
TS20	7	3	44	0	10	11.7	273	200-300
TS21	6	3	0	10	20	11.2	347	100-400
TS24	9	5	39	2	3	22	166	100-200
TS25	8	4	39	0	16	6.6	279	200-300
TS28	8	3	17	8	11	15.7	341	200-500
TS29	13	8	56	0	5	50.8	741	300-2000
TS31	9	4	22	0	19	9.3	387	200-400
TS32	>10	8	17	3	9	30.1	738	400-800 & c. 2kb
TS33	4	2	0	6	3	27.7	366	300-400

* Only amplified sample number 4

TS18, 22, 23, 26, 27 and 30 failed to amplify

TS1 and TS17 generated multiple bands using all panel members, including clones

Table 2.4. Polymorphic markers for population genetic analysis

Ten markers located on all four chromosomes were selected for population genetic studies. The TA references in the 'Closest CDS' column are the designated identification numbers for coding sequences as used by the online genomic resource, GeneDB. 'Copy no.' refers to the copy number of the repeat motif in the C9, genome strain of *T. annulata*. The chromosomal positions of the loci are detailed elsewhere in Figure 2.4.

Table 2.4. Polymorphic markers for population genetic analysis

Name	Chromo -some	Closest CDS	Location of repeated region	Consensus repeat sequence	Repeat motif copy no.	*T. annulata* size range (bp)	*T. lestoquardi* size (bp)	PCR primers 5' – 3'	Annealing Temp. (ºC)
TS5	4	TA10045	exon	GGTTCA	13.8	240 - 318	228	F ctggaacatgaattacttgttcttcc R ggacaccaatgagtgacgtgacag	60
TS6	4	TA11040	intron [c]	TAATTATAGG	13.6	301 - 466[a]	266	F catcctttgacctactgattgtac R cggtagtaccagttaatactgtc	60
TS8	3	TA03940	exon	TATTATTTAATG	11.2	195 - 356	192	F taaacgattaaaatcaagtg R attggaaatggtgaaataatgag	55
TS9	3	TA03885	intergenic [d]	ATT	27.3	338 - 386	378	F aatgtgtggtacaacatcac R gatatggaatcatactagaagttg	58
TS12	3	TA18345	exon	AATACT	10	237 - 376	299	F gatgatagaggaattgatatgac R ggaaatatcacaattaagattc	55
TS15	1	TA20375	exon	AAGATACTAATGGAAGATTAAGTA	6.7	164 - 404	148	F gtacgtaatcttggaaatggtag R gatacaacgttacggagtcagttgg	60
TS16	1	TA20830	intergenic [d]	TAA	35	345 - 439	349	F ccaatgtcaacagtatgatg R gagtaagaagtaccactactg	56
TS20	2	TA13850	intron [c]	ATTATTACTA	11.7	187 - 310	340	F ccttcatgatctacatctgatgc R ggctgaatgggtacctgttc	60
TS25	4	TA08330	intron [c]	ATTATACTATACTATT[b]	6.6	209 - 296	197	F cgccatcagtagtcatctcag R gacgaccataactgggaagtcaac	60
TS31	2	Unknown	non-coding region	AATTTATCCTGAATTATAGA	9.3	203 - 385	No amplification	F gttatcttcttgctattatagc R gtattaaaatctataagattc	50

[a] 3 alleles detected above 500bp, approx 505, 600 and 600 bp

[b] several other repeat motifs identified: TATAC , TATACTATTAT

[c] both primers in exons

[d] single primer in exon

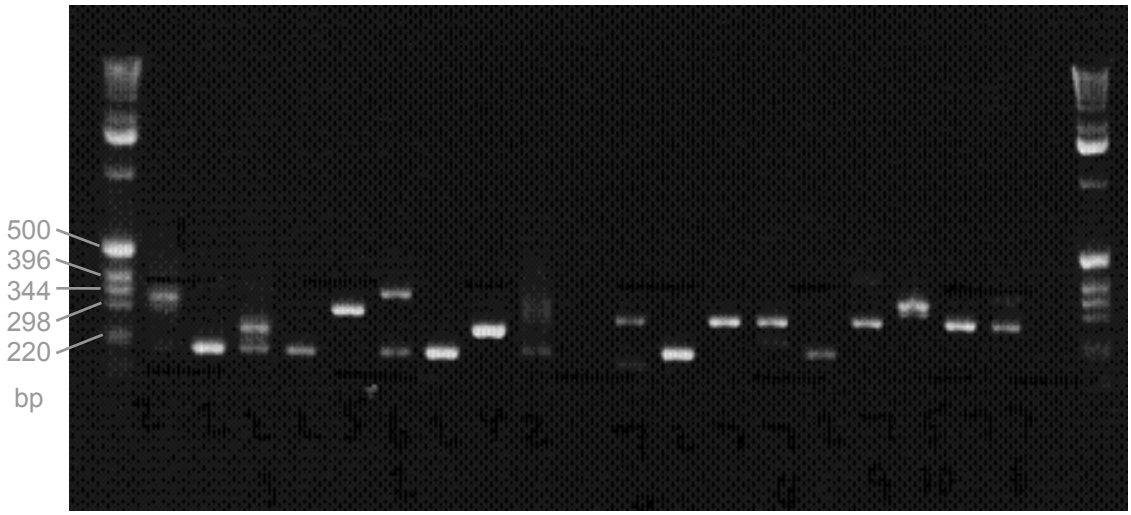
Figure 2.2. Example of agarose gel electrophoresis of marker TS15

This image represents PCR products generated using the TS15 primer set on the panel of DNA samples that was used to screen all of the markers. An aliquot of PCR product from each reaction was loaded into a 2 % agarose gel and separated by electrophoresis. Lanes 1 and 22 contain a 1 kb ladder used as a reference to determine the size of the amplicon in each lane; the sizes of five (known) bands are indicated on the left of the diagram. PCR products vary in size from around 200 to 400 base pairs (bp). Although two bands are evident in the clonal sample in lane 12 (*), this isolate was later determined to be a mixed infection.

Figure 2.3. Example of Genescan™ analysis

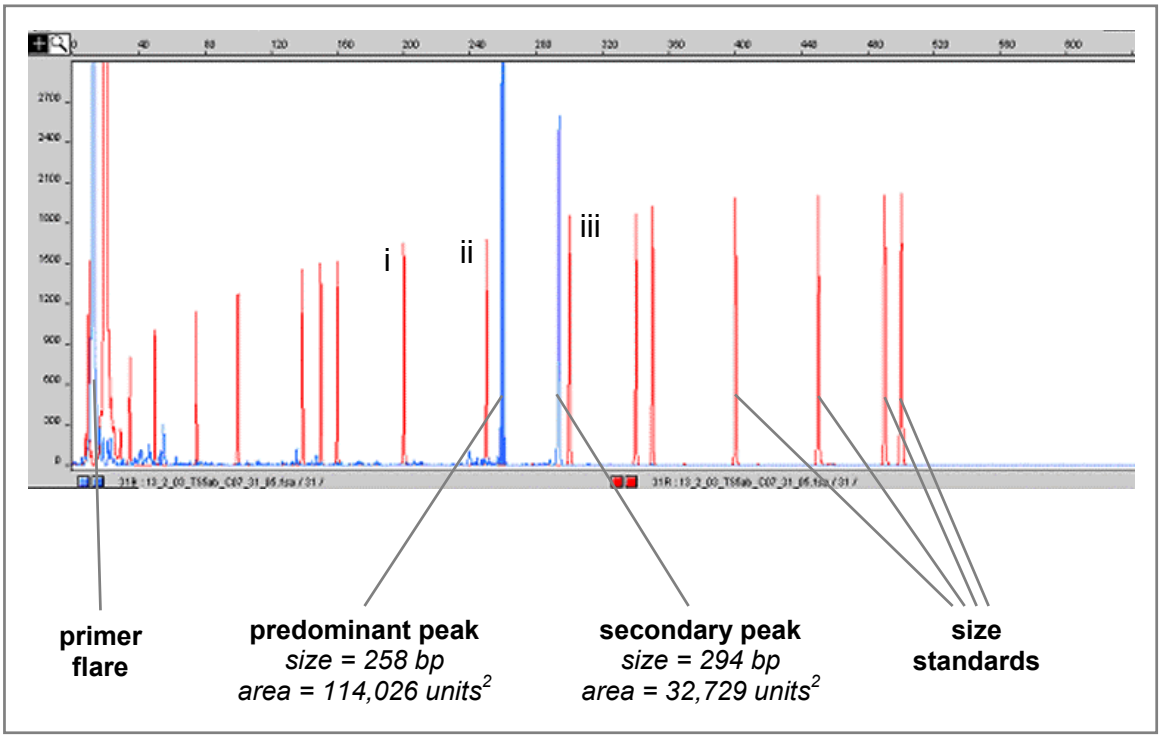
This image represents a trace, or electrophoretogram, generated by Genescan™ analysis and depicts fluorescently-labelled PCR products. This particular trace was produced using the TS5 primer set with a DNA template containing a mixture of two genotypes, representing an early passage of a Tunisian cell line (LBM-23, Ben Miled, 1993). The horizontal axis represents fragment size while the vertical axis represents units of fluorescent intensity. Blue peaks represent labelled PCR product, while red peaks indicate the GS500 size standards: i = 200 bp, ii = 250 bp and iii = 300 bp. The sizes of the alleles represented by the two blue peaks are interpolated from their position with respect to the size standards. The predominant allele, 258 bp is defined as the peak with the greatest area under the curve.

Figure 2.2. Example of agarose gel electrophoresis of marker TS15



Lane	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Contents	*size marker*	field isolate	field isolate	field isolate	cell line	clone	cell line	cell line	cell line	cell line	*T. sergenti*	clone *	clone	clone	cell line	cell line	cell line	cell line	cell line	cell line	-ve control	*size marker*

Figure 2.3. Example of Genescan™ analysis



2.2.4. Data analysis

Similarity comparison of MLGs was undertaken using an allele sharing co-efficient (Bowcock *et al.* 1994) in Excel Micro-satellite Toolkit (Park 2001). Comparison of stocks and isolates containing mixtures of genotypes was achieved using Jaccard's similarity (Jaccard 1908). This was undertaken by creating a character matrix input file, representing the complete allelic profile of every sample. The web-based application, Clustering Calculator, (http://www.biology.ualberta.ca/old_site/jbrzusto/cluster.php) was used to calculate the pair-wise similarity of allelic profiles between all the samples. To illustrate, Jaccard's similarity between samples i and j is represented by the formula –

$$\text{Jaccard's similarity} = \frac{\text{no. of alleles present in both i \& j}}{\text{(no. of alleles present in both i \& j + no. of alleles present in i, absent j + no. of alleles present in j, absent in i)}}$$

For each pair-wise combination, a value between 0 (completely dis-similar) and 1 (identical) was calculated based on the number of alleles that the two profiles share. This program was also used to cluster the data to produce dendrograms, which were visualised using TreeViewX version 0.4 (http://darwin.zoology.gla.ac.uk/%7Erpage/treeviewx/). Bootstrap values for the consensus trees were also calculated by Clustering Calculator using 1,000 pseudo-replications. Population genetic analysis by estimation of F-statistics was performed using Fstat v. 2.9.3.2 (http://www2.unil.ch/popgen/softwares/fstat.htm) and Nei's Genetic Distance (Nei 1978) was calculated by the GDA software application (http://lewis.eeb.uconn.edu/lewishome/gda.html). The null hypothesis of linkage equilibrium was tested using LIAN (http://adenine.biz.fh-weihenstephan.de/lian/), which also calculated the standardised index of association (I_S^A) (Haubold and Hudson 2000), a quantification of linkage equilibrium/disequilibrium. Linkage equilibrium (LE) is characterised by the statistical independence of alleles across all loci under investigation. LIAN tests for this independent assortment by initially determining the number of loci at which each pair of MLGs differs. From the distribution of mismatch values, a variance V_D is calculated. This value is compared to the variance expected for LE, which is termed V_e. The null hypothesis that $V_D = V_e$ is tested by (a) a Monte Carlo computer simulation and (b) a parametric method. The software returns the 95 % confidence limit, L for both methods, which are denoted L_{MC} and L_{PARA} respectively. When V_D is found to be greater than L, the null hypothesis is disproved and linkage disequilibrium is indicated.

2.3. Results

2.3.1. Identification and evaluation of markers

Screening the genome sequence with the tandem repeat finder program identified 3,206 repetitive sequences, as defined by the input parameters. However, many of these repeated sequences either overlapped partially or completely, with variant motif forms identified at the same locus. A filtration process was used to identify a manageable subset of loci, which could be tested using the panel of stocks and isolates. This included discarding repeat regions greater than 500 bp in length and also those possessing insufficient flanking sequence to design primers. Remaining sequences were ranked, based on the fidelity of the repeat within each region (> 70 % fidelity) and the number of repeats. Thirty-three top ranking loci were identified (Table 2.3.), comprising 10 micro-satellites (motif size 2 - 6 bp) and 23 mini-satellites (motif size 9 - 54 bp). The high number of repeat regions identified (approximately 1 per 3 kb, on average) was surprising because, although such loci are known to be widely distributed over eukaryotic genomes, the *T. annulata* genome has been shown to be 'gene dense' with the amount of non-coding sequence normally associated with satellite repeat regions being relatively small (Pain *et al.* 2005). Primers were designed to unique sequence flanking each of the 33 satellite loci and used to amplify DNA from the panel of stocks detailed in Table 2.2. Of the 27 repeat regions amplified by the primer pairs, 7 were located in exons, 7 were in introns, with primers located in the flanking exons and 7 in intergenic regions, while one (TS21) was found to straddle an intron and an exon. The 5 remaining repeat regions were in contigs without significant open reading frames. The genes associated with or flanking the markers are generally annotated as hypothetical proteins of which more than 90 % have orthologues present in the *T. parva* genome (Gardner *et al.* 2005).

The initial screen was designed to characterise the loci and determine whether each marker represented a single locus, produced an amplicon of the predicted size and was species specific, polymorphic and able to amplify DNA from all samples in the panel of stocks. To illustrate such an appropriate marker, an agarose gel displaying the amplification products of the TS15 primer set is presented in Figure 2.2. These primers were shown to be able to PCR amplify all 19 DNA templates and a considerable degree of size polymorphism was evident across the panel. Single alleles were identified in the lanes representing clones (6, 13 and 14). Multiple bands were associated with field isolates (lanes 2 and 4) and cell line extracts (lanes 7, 10, 15, 17, 18 and 20) indicating the presence of a mixture of genotypes. The DNA preparation 'Bat cl4 clone 5' in lane 12 displayed

two alleles both at this locus and six of the other nine loci analysed. Therefore, this stock was deemed to represents a mixture of haploid genotypes rather than being a true clone with two TS15 loci. The *T. sergenti* DNA template failed to amplify, suggesting the TS15 primers may not co-amplify DNA from other *Theileria* species. Each of the 31 sets of primers were used to PCR amplify the panel of stocks and isolates and the results are summarised in Table 2.3. Since *T. annulata* is haploid, markers TS1 and TS17 were discarded as several PCR products were detected in samples known to be clones. For all the markers, the allele sizes amplified from the cloned C9 genome sequence stock (as determined by agarose gel electrophoresis) were consistent with sizes predicted from the genome sequence.

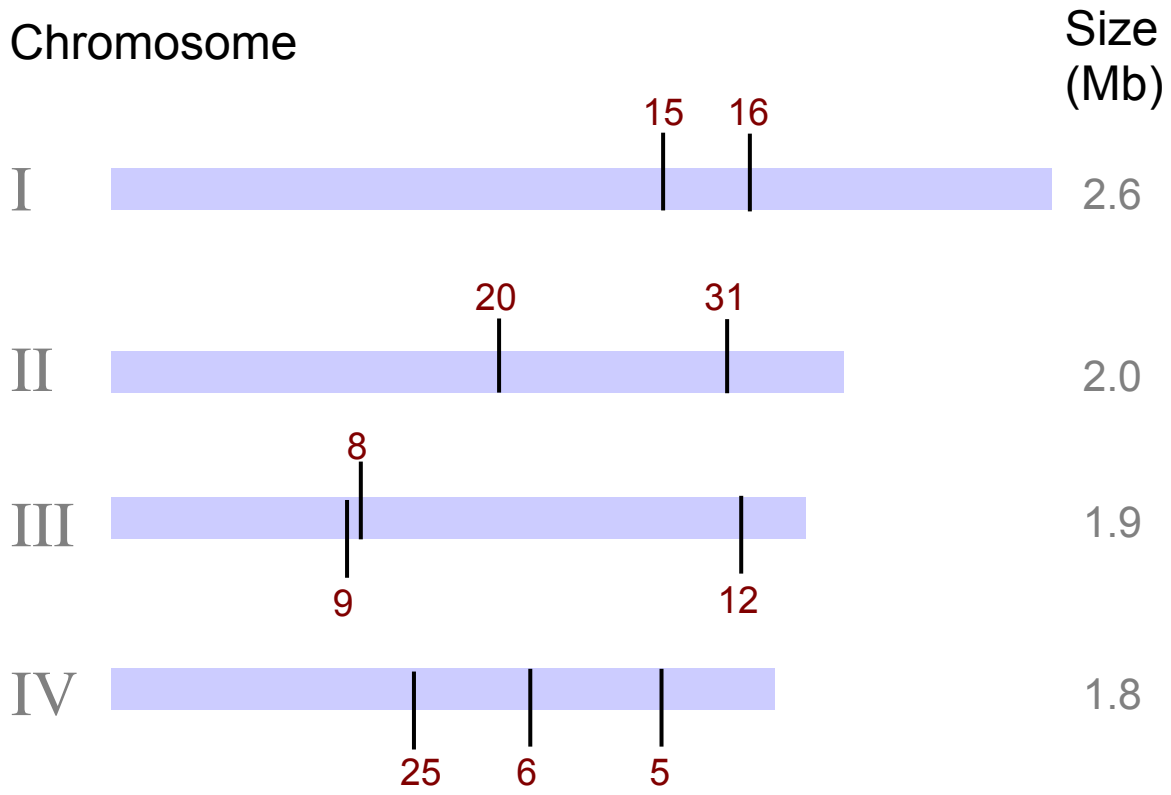
As sequencer-based analysis was to be undertaken, amplicons greater than 500 bp (TS2, TS29 and TS32, see Table 2.3.) were deemed unsuitable for further study as they were outside the reference range of the GS500 size-standards. Six of the 33 primer pairs failed to amplify any product. Primers which failed to amplify two or more of the *T. annulata* DNA samples were not analysed further, leaving a panel of four micro- (TS5, 9, 12 and 16) and six mini-satellites (TS6, 8, 15, 20, 25 and 31) for further study. The characteristics of these loci are summarised in Table 2.4., with the copy number of each repeat found in the genome strain. An example electrophoretogram generated by Genescan™ analysis is shown in Figure 2.3., representing the amplification products of a Tunisian cell line DNA preparation using the TS5 primer set. This example shows two clearly defined blue peaks, the positions of which are interpolated from the distribution of the size standards, the largest of which is 500 bp. Two alleles of 258 bp and 294 bp are identified, with the former being most abundant by virtue of its greater area under the curve. The ten selected markers all showed polymorphism both between stocks isolated from different geographical regions as well as between stocks isolated from the same region (Tunisian samples). Each marker demonstrated the ability to detect more than one allele in some of the samples, while some (TS15, 16 and 20) were able to identify a high proportion (> 40 %) of stocks and isolates containing multiple genotypes (Table 2.3.). The ten markers are distributed over all four chromosomes as illustrated in Figure 2.4., with only two loci physically closely linked (< 100 kb) - TS8 and TS9, which are separated by 38 kb.

A non-transforming *Theileria* species, *T. sergenti*, outside the *T. annulata*, *T. lestoquardi*, *T. parva* and *T. taurotragi* cluster, as inferred from a phylogenetic tree based on 18S RNA (Schnittger *et al.* 2003) was initially selected to test for species specificity. The aim was to exclude markers with a risk of cross species amplification, without discarding potentially

Figure 2.4. Distribution of the ten selected markers across the genome

The ten markers selected for further analysis are distributed over the four chromosomes of the *T. annulata* genome and lie outwith the telomeres. The numbers in red correspond to the TS numbers used to denote each marker. Only two loci, TS8 and TS9, are physically closely linked and are separated by approximately 38 kb of sequence. The nearest annotated coding sequence (CDS) to each locus is presented elsewhere (Table 2.4.).

Figure 2.4. Distribution of the ten selected markers across the genome



informative ones, which may still amplify closely related species. *T. sergenti* DNA failed to amplify with any of the sets of primers. The markers were further tested against a panel of DNA from another five different *Theileria* species (detailed in Section 2.2.1.). The primers failed to amplify DNA from any of these species, apart from *T. lestoquardi*. With the exception of TS31, all selected pairs of primers generated amplicons using DNA from *T. lestoquardi*. Only one allelic product was detected for each marker, consistent with the *T. lestoquardi* stock being a clone. Allele sizes for five of these markers were smaller than those obtained for any *T. annulata* allele, with only one of the nine markers generating a product with a size greater than any identified for *T. annulata*. The other three PCR products had sizes similar to those identified in *T. annulata*. These results would not confuse the analysis of *T. annulata* in cattle, since *T. lestoquardi* is restricted to small ruminants.

2.3.2. Diversity of markers used for genetic analysis

The ten markers described in Table 2.4. and Figure 2.4. were used to genotype DNA samples representing isolates from Tunisia, Turkey and other countries (Table 2.1.), using a DNA sequencer to determine the size of each amplicon. As illustrated in Table 2.5., all markers were found to be polymorphic for both the Tunisian and Turkish samples. Across all 72 samples, the number of alleles ranged from 9, for micro-satellite TS5, to 29, for markers TS8, TS31 (mini-satellites) and TS12 (a micro-satellite). The highly polymorphic TS8 identified the maximum number of alleles by any one marker in a single DNA preparation, discriminating ten alleles in one of the Turkish stocks. Primers representing markers TS6 and TS25 failed to amplify a PCR product from the Spanish stock, whereas only TS16 failed to amplify from two stocks in the Tunisian population. Since all other primer sets amplified from these three samples, failure to generate PCR product was attributed to polymorphism at the primer sites in these stocks. The average number of alleles for each marker was 20.6 and there was no significant difference between micro-satellites and mini-satellites in the number of alleles identified. However, two broad patterns of diversity were evident: firstly, alleles differing in size by the unit repeat motif length, with no intermediate sizes (TS5, TS8, TS15 and TS25) and secondly, a more continuous spectrum of sizes (TS6, TS9, TS12, TS16, TS20 and TS31) with both micro- and mini-satellites represented in each class. These two patterns of variation are illustrated in Figure 2.5. for markers TS5 and TS20 using the data from all 72 samples in the study. Mutations in micro- and mini-satellites are thought to occur by replication slippage or unequal crossing over during meiosis, resulting in changes in copy number (Debrauwere *et al.* 1997). Step-wise mutation and infinite allele models have been proposed to describe

Table 2.5. Allelic variation in Tunisian and Turkish populations

All 72 DNA samples were genotyped using the panel of ten markers. The minimum and maximum number of alleles detected at each locus within a sample was determined across all ten markers for each sample that amplified. The number of alleles represented in each population was calculated, taking into account the most abundant and all the minor alleles present in each sample from that country. Gene diversity was calculated for each marker and is equivalent to estimated heterozygosity.

Table 2.5. Allelic variation in Tunisian and Turkish populations

		n	TS5	TS6	TS8	TS9	TS12	TS15	TS16	TS20	TS25	TS31
Number of alleles within each sample	**Minimum**	**72**	1	1	1	1	1	1	1	1	1	1
	Maximum	**72**	6	7	10	8	9	7	4	6	7	8
	No amplification	**72**	0	1	0	0	0	0	2	0	1	0
Number of alleles in population *(both primary & secondary)*	**Tunisia**	**51**	8	16	24	22	18	8	12	13	7	18
	Turkey	**13**	6	10	8	9	10	7	7	10	7	10
	Other countries	**8**	6	7	5	6	7	4	7	6	5	8
	Overall	**72**	9	24	29	25	29	10	19	20	12	29
Gene diversity	**Tunisia**	**51**	0.813	0.896	0.952	0.954	0.925	0.804	0.854	0.890	0.707	0.930
	Turkey	**13**	0.859	0.962	0.923	0.936	0.949	0.872	0.897	0.949	0.795	0.974

n = number of isolates

Figure 2.5. Variation of allele size-intervals between markers

Two broad patterns of diversity were defined for the ten selected markers and an example of each is presented opposite. In each case, alleles are ranked in order of increasing length from left to right to highlight the differences between neighbouring alleles and to indicate the frequency of each allele in the overall dataset.

(i) Regular intervals between allele sizes (TS5)

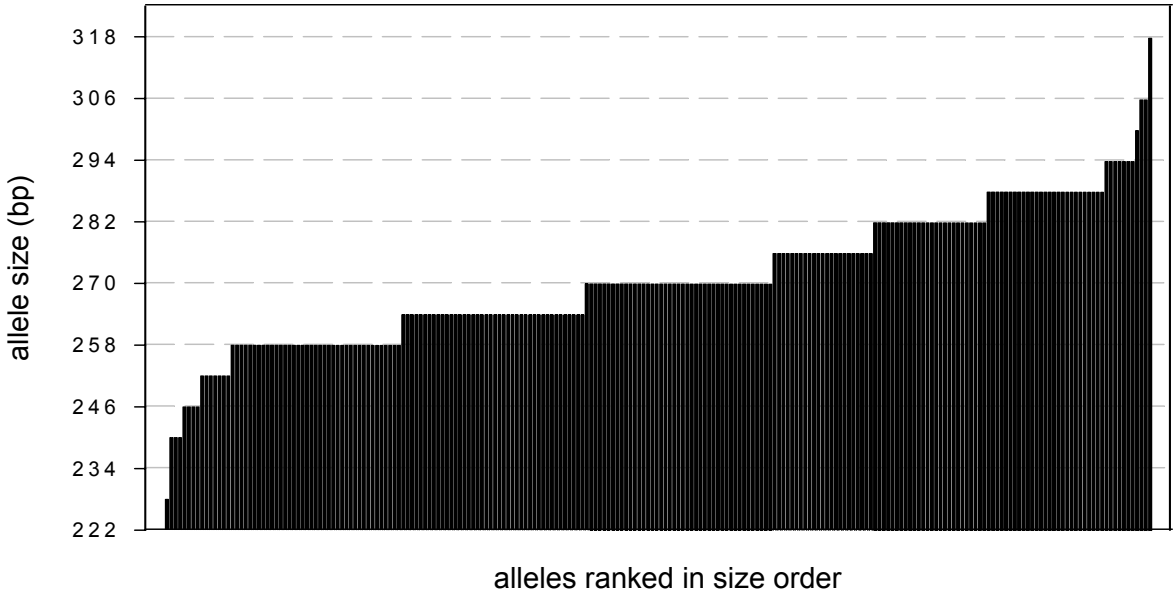
Alleles differ in length by six base pairs, corresponding to the repeat motif, GGTTCA. A limited number of distinct alleles are present at a relatively high frequency in the population. This pattern of distribution is consistent with a 'step-wise' mutation mechanism, often considered a classical characteristic of micro-satellite loci.

(ii) Irregular intervals between allele sizes (TS20)

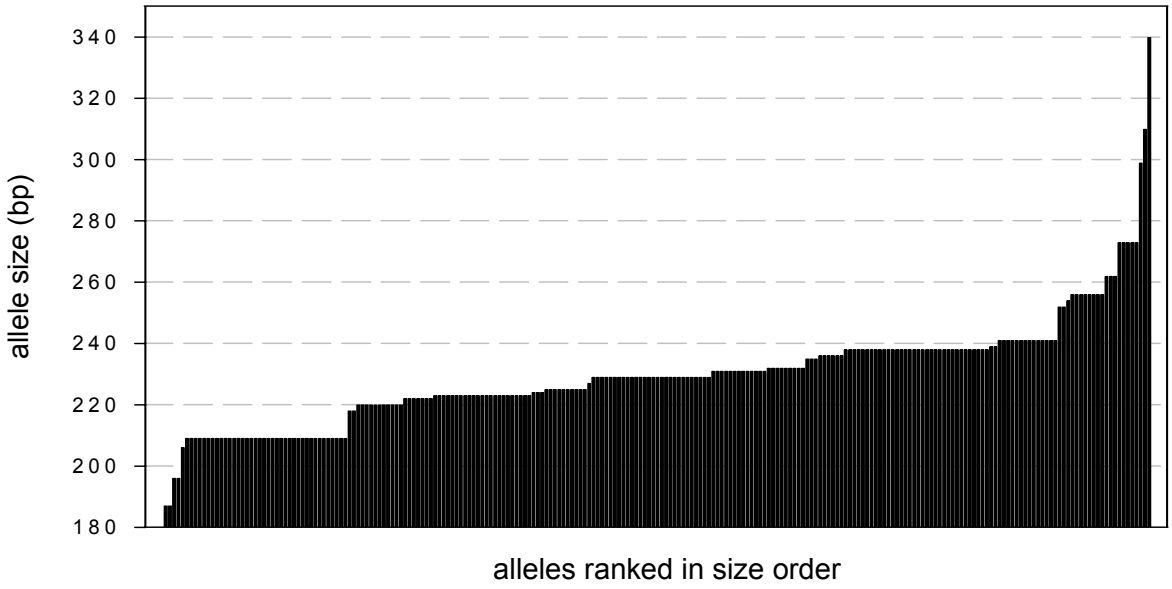
Alleles differ by a variable length, with a continuous gradation exhibited in parts of the range with a higher number of distinct alleles observed at a relatively lower frequency in the population. A 10 bp mini-satellite motif, ATTATTACTA, was identified at this locus and much of the variation is attributed to the internal ATT repeat. This pattern of distribution is more consistent with the 'infinite allele model', a mechanism often associated with mini-satellite loci, whereby every length mutation in the locus results in a novel allele being generated.

Figure 2.5. Variation of allele size-intervals between markers

(i) Regular intervals between allele sizes (TS5)



(ii) Irregular intervals between allele sizes (TS20)



the resulting allelic variation, which can explain the patterns of diversity exhibited by these markers (Zhivotovsky and Feldman 1995). Although sequencing a number of alleles from each locus would be necessary for confirmation of the mutational mechanism, the data based on size (Figure 2.5.) suggest that mutation in the first group of markers occurs in a step-wise fashion whereas, in the second group, a more complex mechanism is operating. An almost continuous gradient of allele sizes in the range of 220 – 240 bp was evident for marker TS20, but larger changes in allele size outside this range occur suggesting that a variety of mechanisms may influence diversity. The pattern of allele size variation did not relate to whether the locus was a micro-satellite or a mini-satellite. However, three of the four markers that behaved in an apparent 'step-wise' manner (TS5, TS8 and TS24) had their repeat region located in coding sequences. Their motif lengths (6, 12 and 24) are multiples of three, which would allow the maintenance of the open reading frame (ORF) in which they lie without introducing frame shift mutations. No correlation was observed between repeat length and the number of alleles at a locus, however repeat length positively correlated with the PCR product size range ($r^2 = 0.51$, $p < 0.01$). That is to say, larger PCR products were associated with longer repeat motifs, although it must be appreciated that the maximum allele size was constrained, since one of the criteria for selecting the markers was that the largest allele did not exceed 500 bp in length.

2.3.3. Genotyping of cell lines, piroplasms and clones

A mixture of alleles at several loci was detected in more than 90 % of cell lines derived from field samples, with an average of between three and four alleles in these samples. To confirm that the multiple allelic phenotype of cell line DNA reflected a mixture of haploid parasite genotypes, a series of preparations of the Ankara isolate and its derivatives were analysed. A piroplasm extract, a cell line and seven cloned cell lines were genotyped at all ten loci, the results of which are presented in Table 2.6. The piroplasm extract was the most diverse, with as many as six alleles identified by TS31. The cell line showed an intermediate level of diversity and displayed a maximum of two alleles at each locus, while all the clones possessed a single allele at each locus, confirming the presence of a single haploid genotype. Two clonal genotypes were identified, the alleles of which are highlighted in red and blue. Moreover, these genotypes shared only two alleles, which have been highlighted in green. The major MLG generated from the A_2 cell line was identical to one of the two clonal MLGs, while the alleles representing the other clone were all identified as secondary components. The MLG constructed from the most abundant alleles at each locus in the piroplasm extract did not match either the MLG of the A_2 cell line or any of the derived clones. This implied that there were further genotypes within the

Table 2.6. Genotyping of the Ankara isolate and derived clones

DNA samples representing one piroplasm preparation, one cell-line and seven clones all derived from the Ankara strain of *T. annulata* were genotyped at all ten loci. Allele sizes in base pairs were recorded with minor alleles denoted in parenthesis. Coloured numbers represent alleles found in either or both of the two clonal genotypes, i.e. **green** numbers represent alleles which are common to both clones, while **red** and **blue** numbers correspond to alleles present in one clone and absent in the other.

Table 2.6. Genotyping of the Ankara isolate and derived clones

DNA sample	Type	Marker / alleles (bp)									
		TS5	TS6	TS8	TS9	TS12	TS15	TS16	TS20	TS25	TS31
Ankara	piroplasm isolate	**282** (264)	**389** (**396**)	**305** (**284**) (308) (250)	352 (**360**) (362) (**364**)	**267** (258) (292)	308 (**188**) (**284**) (262)	**349**	252 (**223**) (225) (232) (**273**)	215 (**218**) (222) (231) (**280**)	265 (236) (256) (286) (**313**) (**385**)
Ankara A_2	cell line stock	**282**	**396** (**389**)	**284** (**305**)	**360** (**364**)	**267**	**188** (**284**)	**349** (**353**)	**223** (**273**)	**218** (**280**)	**313** (**385**)
A_2 C9	clone	**282**	**389**	**305**	**364**	**267**	**284**	**353**	**273**	**280**	**385**
A_2 B4-1	clone	**282**	**389**	**305**	**364**	**267**	**284**	**353**	**273**	**280**	**385**
A_2 D7	clone	**282**	**389**	**305**	**364**	**267**	**284**	**353**	**273**	**280**	**385**
A_2 D3	clone	**282**	**396**	**284**	**360**	**267**	**188**	**349**	**223**	**218**	**313**
A_2 E3	clone	**282**	**396**	**284**	**360**	**267**	**188**	**349**	**223**	**218**	**313**
A_2 B2-1	clone	**282**	**396**	**284**	**360**	**267**	**188**	**349**	**223**	**218**	**313**
A_2 J1-1	clone	**282**	**396**	**284**	**360**	**267**	**188**	**349**	**223**	**218**	**313**

Ankara isolate that were not represented by the macroschizont infected A_2 cell line. Alternatively, it is possible that the Ankara piroplasm predominant MLG was not truly representative of a single parasite genotype. Such highly heterogeneous DNA preparations may be predisposed to generating erroneous genotyping data. Using a mixed DNA template, heteroduplexes may potentially form between non-identical PCR products resulting in the artefactual creation of novel PCR species. However, in this study, fluorescently labelled PCR products were denatured before Genescan™ assay, hence disassociating any heteroduplexes present in the mixture to produce single-stranded DNA. The possibility of *in vitro* recombination between unlike alleles during amplification of target DNA containing mixed genotypes has been highlighted in *P. falciparum* (Tanabe *et al.* 2002). In the case of the Ankara isolate, the extensive heterogeneity encountered at seven of the ten loci in the piroplasm preparation revealed alleles previously identified in other isolates in the population. This suggested that a high proportion of these secondary alleles were genuine.

These data raise the question of whether low passage cell lines derived from infected animals necessarily reflect the most abundant genotype in the host. To investigate the relationship between homologous preparations from a single animal *in vitro* and *in vivo*, a panel of matching piroplasm and cell line DNA preparations was analysed. The mean number of alleles at each locus was determined and is shown in Figure 2.6., with error bars indicating the standard error. In three out of six cases the piroplasm extracts showed a considerably higher level of diversity than the cell lines (paired t-test, $p < 0.05$). The number of alleles shared by each pair of MLGs (based on the most abundant allele at each locus) is shown on the lower part of Figure 2.6. The lowest identity (5 / 10) was shown in the highly diverse 9A and 9B, while the highest identity (10 / 10 and 9 / 10) was shown in stocks 10 and 13, where the lowest amount of piroplasm diversity was identified. This may be interpreted by two hypotheses that are not mutually exclusive. The first hypothesis is that with increasing heterogeneity of the parasite population in the host, the chances decrease of the most abundant *in vivo* sample establishing as the predominant genotype *in vitro*. This may be explained by either *in vitro* selection or it may alternatively be simply a stochastic effect, which may be investigated by establishing and genotyping more cell lines. A second hypothesis is that with increasing parasite heterogeneity, the constructed MLG for the *in vivo* sample may fail to incorporate alleles of the most abundant *in vivo* genotype, with artefactual alleles being contributed by secondary components in the mixture. These hypotheses are discussed in Section 2.4.5.

Figure 2.6. Multiplicity of infection in cell lines and homologous piroplasm extracts

A panel of homologous piroplasm and cell line DNA preparations, each from a single animal, was analysed, representing six matched pairs of *T. annulata* samples originally isolated in Tunisia. The mean number of alleles at each of the ten loci is presented with error bars indicating the standard error across all ten loci. Multilocus genotypes (MLGs) were created using the most abundant allele detected at each locus in each sample and the number of alleles shared by each pair of MLGs is indicated beneath the chart. Piroplasm preparations displayed more heterogeneity than cell line preparations, particularly those representing the most heterogeneous isolates, however, such isolates shared a low number of alleles between the piroplasm DNA and the cell line DNA. A paired t-test was performed on the number of alleles identified over all ten loci, comparing the six pairs of homologous samples in turn.

9A: $p = 0.022$

9B: $p = 0.070$

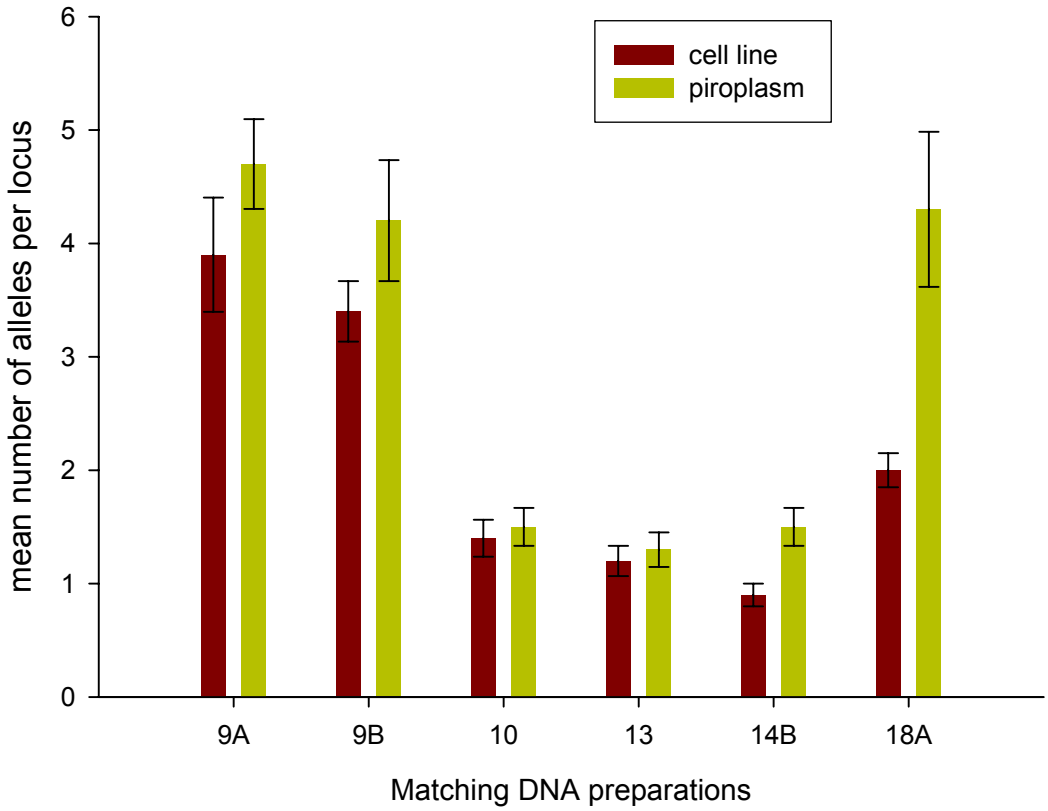
10: $p = 0.343$

13: $p = 0.343$

14B: $p = 0.005$

18A: $p = 0.011$

Figure 2.6. Multiplicity of infection in cell lines and homologous piroplasm extracts



MLG shared alleles	**5 / 10**	**5 / 10**	**10 / 10**	**9 / 10**	**6 / 10**	**7 / 10**

As described in Section 2.2.3, multilocus genotypes (MLGs) were constructed using the predominant allele present at each locus and used to generate the dataset for the population genetic analysis (Table 2.7.). Each stock had a unique MLG, with the greatest identity between any two samples being six out of ten allelic markers. Predominant allele frequencies were determined for Tunisian and Turkish populations and a high level of diversity was observed within each country at all loci tested, as illustrated in Figure 2.7. A spectrum of variation was observed across the markers when comparing the two populations. That is to say, at one extreme, limited differentiation was detected where there were few private alleles specific to each population and similar allele frequencies were observed in each country (Figure 2.7., TS5 and TS15). For TS5, only one private allele (246 bp) was present in the Tunisian population and similar frequencies were encountered for five of the other seven alleles (258, 264, 276, 282 and 288 bp). At the other extreme, markers displayed a high level of divergence where the most frequent alleles were private and common alleles showed large differences in allele frequencies (Figure 2.7., TS6, TS16 and TS25). In the case of TS6, the three most frequent alleles in the Tunisian population were not found in Turkey and, of the 21 (predominant) alleles identified for this marker, only three were common to both countries. In general, across the markers more private alleles were detected in Tunisia, probably reflecting the larger sample size from this population.

2.3.4. Genetic analysis of *T. annulata* populations

Principal MLGs representing stocks from the populations of Eastern, Western and Central Tunisia, Northern and South-western Turkey and Sudan were analysed using standard techniques to measure heterozygosity, linkage disequilibrium and population differentiation. Since *T. annulata* is haploid and heterozygosity cannot be observed, the estimated heterozygosity (H_e) was calculated from the predominant allele data set. Heterozygosity within each regional population was generally high, ranging from 0.86 within Central Tunisia to 0.93 in South-western Turkey (Table 2.8.). The number of genotypes present in an individual sample was estimated using the number of alleles identified by the most polymorphic markers for that sample, with a mean value calculated over each population. Mixtures of genotypes in individual stocks were common in each of the regional populations. The mean number of alleles per locus ranged from 3.00 to 3.43 for the Tunisian and Sudanese populations, suggesting on average each sample contained at least three distinct genotypes. After three laboratory-generated clones were excluded from the Turkish population, a lower value of 2.40 indicated that on average between two and three genotypes were present in each stock.

Table 2.7. Multilocus genotypes used in population genetic analysis

Samples representing 64 isolates derived from two Turkish provinces and three regions in Tunisia were used for formal population analysis. Four Sudanese samples were also included for a limited portion of the analysis. A multilocus genotype (MLG) was generated for each sample, representing the most abundant allele detected at each of the ten loci. Allele sizes are indicated in base pairs (bp).

Table 2.7. Multilocus genotypes used in population genetic analysis

Population	TS5	TS6	TS8	TS9	TS12	TS15	TS16	TS20	TS25	TS31
Ankara province, Turkey	282	389	305	364	267	284	353	273	280	385
	258	362	264	375	317	188	353	223	222	273
	258	401	301	358	270	238	349	222	218	292
	264	389	305	352	282	356	350	236	218	286
	252	392	264	364	276	248	367	196	215	256
	282	396	284	360	268	188	349	223	218	314
Aydın province, Turkey	264	367	273	358	296	238	345	223	235	248
	258	392	308	374	282	188	349	225	218	256
	258	403	284	357	288	238	379	220	210	293
	288	383	250	361	274	262	363	239	235	274
	252	452	333	360	260	262	367	236	218	203
	276	401	308	360	274	308	350	209	218	287
	252	466	305	369	282	238	350	256	213	292
Central Tunisia	282	401	240	385	290	212	379	229	213	293
	258	353	311	364	290	284	355	229	213	301
	270	443	221	372	250	308	350	223	213	292
	270	383	298	361	376	262	350	209	218	292
	270	383	278	342	270	284	351	231	250	292
	282	416	237	367	250	164	349	238	213	335
	282	416	298	347	276	188	351	223	213	303
Eastern Tunisia	246	416	287	369	270	188	349	209	213	303
	276	367	298	355	261	262	349	238	213	292
	270	362	295	362	253	188	352	238	250	291
	258	365	208	367	298	238	370	209	241	302
	270	392	264	364	267	188	349	209	213	333
	258	416	243	355	250	238	351	256	250	311
	270	365	324	361	284	212	354	222	241	256
	264	403	318	342	267	188	355	209	235	302
	270	373	224	350	276	164	NA	220	213	291
	270	416	237	342	267	262	355	209	213	302
	258	383	284	367	250	308	355	238	241	333
	264	416	275	357	253	262	355	229	241	291
	264	373	261	354	317	262	349	229	213	256
	264	373	298	355	250	308	355	229	213	342
	264	416	247	371	278	188	354	229	213	273
	270	386	264	362	317	164	350	229	213	292
	258	403	270	350	253	188	355	220	241	301
	264	416	224	361	267	212	355	229	213	265
	282	416	224	353	276	284	355	241	213	285
	282	353	237	357	292	238	355	231	213	314
	258	365	275	361	298	238	370	232	250	302
	288	416	224	374	284	188	355	238	210	292
	270	443	258	364	330	262	349	224	241	305
	264	383	275	355	237	188	352	223	213	290
	264	392	234	353	273	188	355	225	213	256
	270	365	224	373	330	334	353	256	213	292
	258	360	284	372	270	188	349	209	213	285
	288	362	247	361	259	262	355	229	215	301
	252	367	264	352	276	262	355	241	250	286
	270	360	298	373	253	188	379	238	218	265
	258	373	237	350	267	262	352	223	213	303
	276	373	237	364	345	262	352	220	250	291
	270	403	237	358	250	308	NA	238	213	333
	282	373	250	354	250	262	409	241	218	302
Western Tunisia	258	403	305	352	253	164	350	238	250	302
	264	416	250	355	270	262	364	229	241	292
	270	420	264	362	273	262	355	220	250	256
	264	373	264	366	276	284	350	231	241	286
	276	365	278	344	253	262	352	231	250	256
	264	416	237	352	250	262	354	229	250	292
	264	305	301	363	317	262	409	223	215	291
	276	423	258	361	250	262	355	209	250	293
	270	416	234	350	267	262	409	241	250	303
	258	365	240	355	267	308	394	187	213	286
Sudan	258	323	261	371	251	212	376	225	209	255
	240	423	231	374	276	238	353	220	218	338
	246	392	261	372	296	308	348	220	265	269
	258	348	231	372	299	212	356	222	218	292

Figure 2.7. Tunisian and Turkish allele frequencies

The frequency of each allele in the Tunisian and Turkish population was determined as a percentage (%) of the total for each marker using the most abundant allele in the 51 Tunisian and 13 Turkish samples described in Table 2.1. These histograms were directly generated from the multilocus genotype data contained in Table 2.7.

Figure 2.7. Tunisian and Turkish allele frequencies

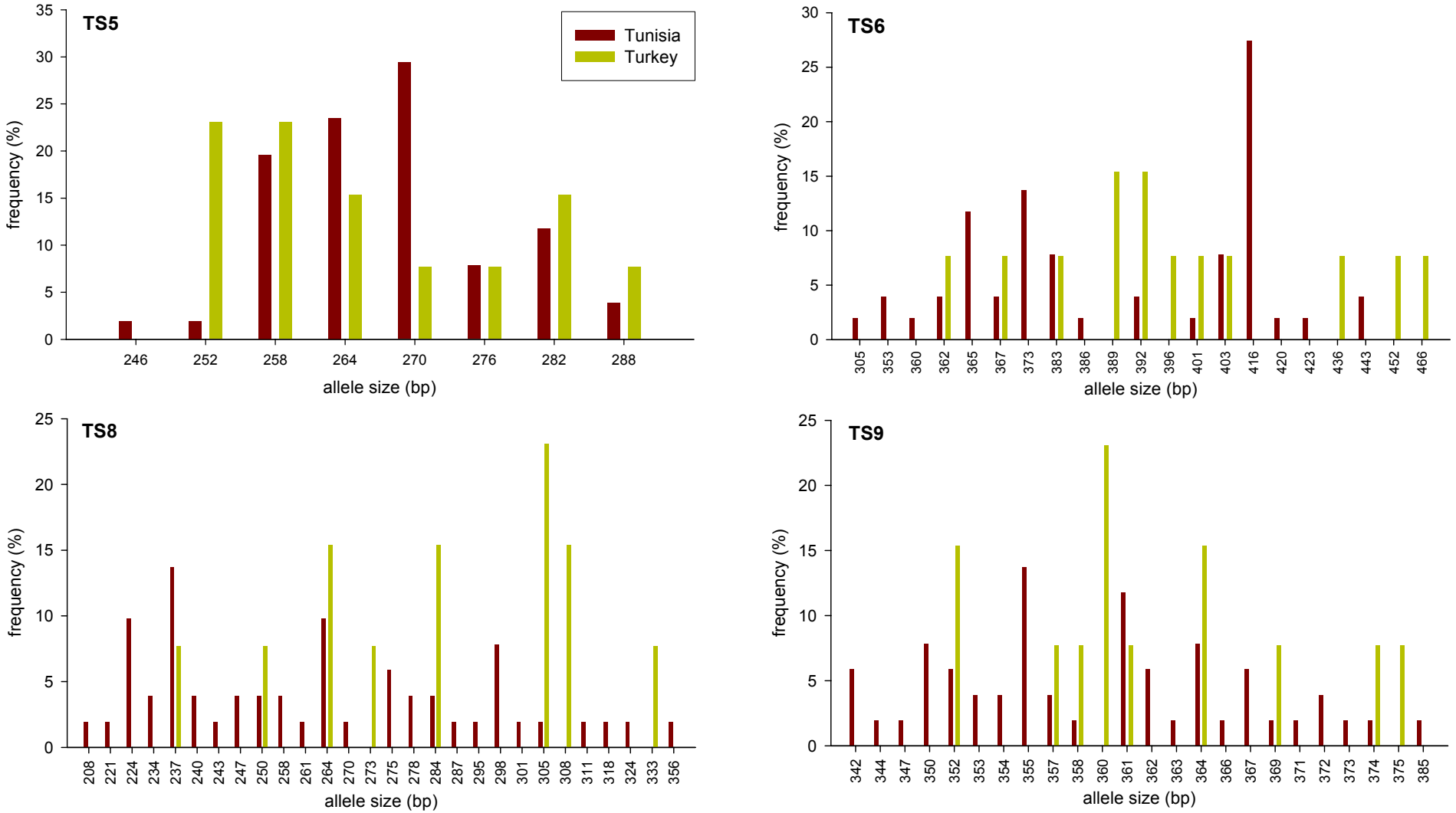
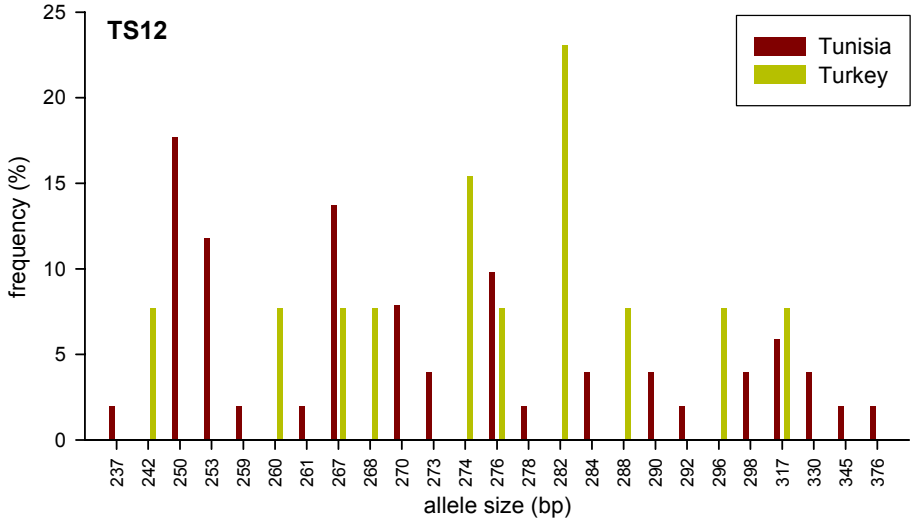
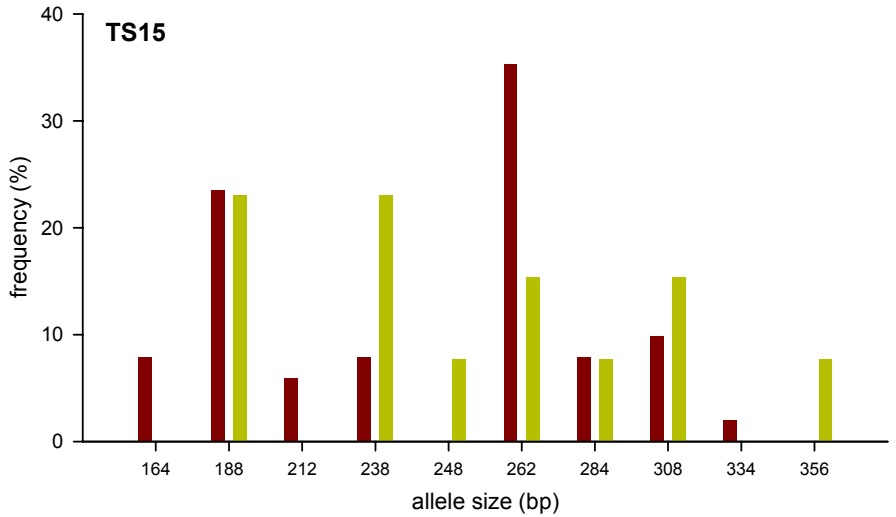
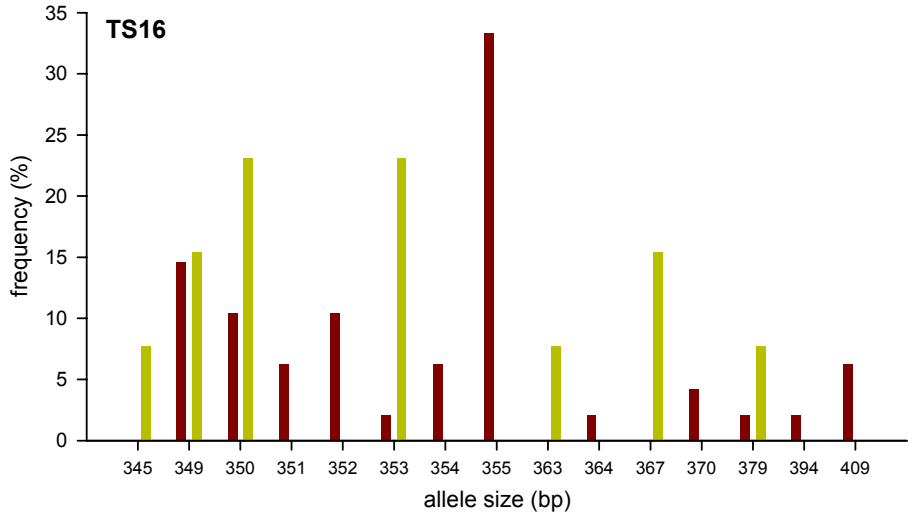


Figure 2.7. Tunisian and Turkish allele frequencies (continued)







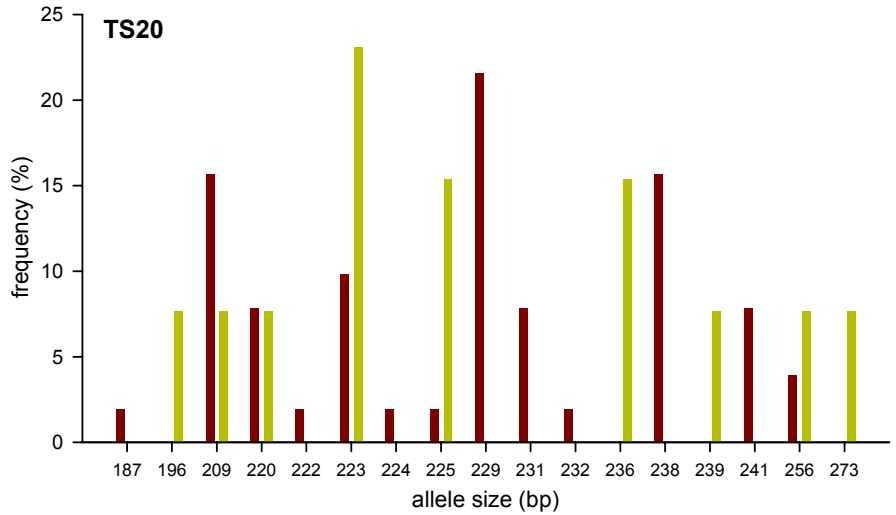
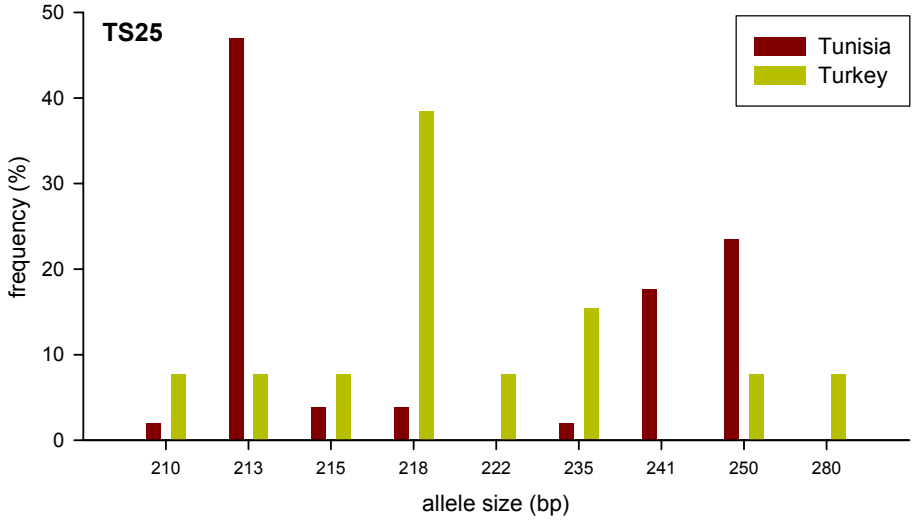


Figure 2.7. Tunisian and Turkish allele frequencies (continued)



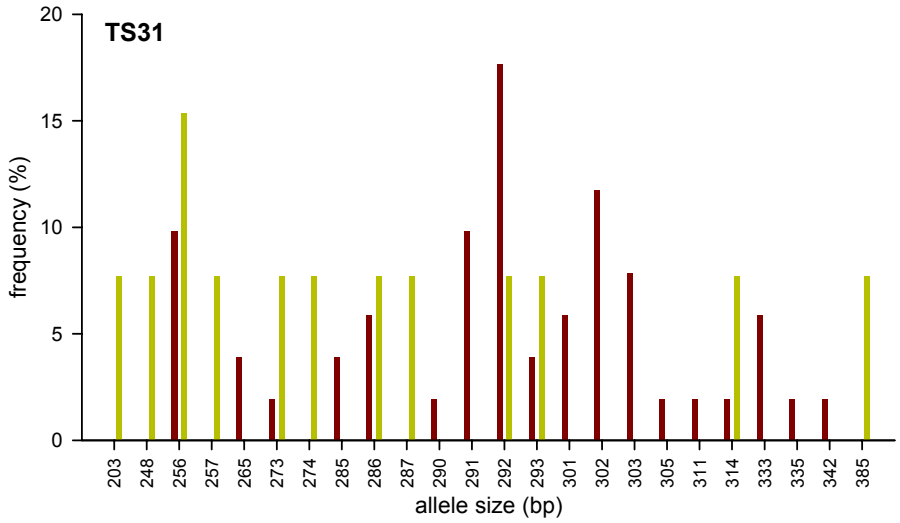


Table 2.8. Population genetic analysis

Standard population tests were conducted on parasite samples representing (a) different countries and (b) areas within Tunisia and Turkey. The 'mean number of genotypes per sample' was calculated as the mean value for the number of alleles detected at each of the ten loci. Structured combinations of populations were pooled to test for linkage disequilibrium. Variance of mismatch values (V_D) were compared to values of L (the upper confidence limits of Monte Carlo simulations and parametric tests), and where $V_D > L$, linkage disequilibrium (LD) was indicated. When $L > V_D$ the null hypothesis of linkage equilibrium (LE) was not disproved. G_{ST}' and θ were used to estimate population differentiation with the standard error for θ calculated in order to assess variance between loci.

Table 2.8. Population genetic analysis

Comparison	n	H_e	Mean no. of genotypes per sample	I^S_A	V_D	L_{MC}	L_{PARA}	Linkage	F_{ST}		
									G_{ST}'	θ	θ SE
Between countries											
Turkey	13	0.912	2.40 *	0.0120	0.945	0.896 *(p = 0.010)*	0.896 *(p = 0.000)*	LD	0.056	0.058	0.016
Tunisia	51	0.872	3.22								
Sudan	4	0.883	3.00								
Turkey	13	0.912	2.40 *	0.0095	0.978	0.950 *(p = 0.040)*	0.950 *(p = 0.000)*	LD	0.044	0.045	0.018
Tunisia	51	0.872	3.22								
Within Tunisia											
East	34	0.867	3.18	-0.0009	1.046	1.127 *(p = 0.550)*	1.128 *(p = 1.000)*	LE	0.023	0.016	0.015
West	10	0.858	3.20								
Central	7	0.862	3.43								
Within Turkey											
Ankara province	6	0.913	2.33 *	0.0001	0.779	0.935 *(p = 0.550)*	0.971 *(p = 0.996)*	LE	-0.021	-0.021	0.008
Aydın province	7	0.929	2.43								

n = number of samples, H_e = estimated heterozygosity, I^S_A = standard index of association, V_D = mismatch variance (linkage analysis),

L_{MC} and L_{PARA} = upper 95 % confidence limits of Monte Carlo simulation and parametric tests respectively (linkage analysis),

G_{ST}' and θ = estimators of F_{ST} (a measurement of differentiation), SE = standard error, LD = linkage disequilibrium, LE = linkage equilibrium, * excludes three laboratory clones

In order to test whether the population of parasites from the three countries comprised a single population with a high level of genetic exchange, the level of linkage equilibrium between pairs of loci was measured using the standard index of association (I^S_A) calculated from MLG data. If there is limited or no association between alleles at different loci, a value close to zero is obtained, whereas if association is detected a positive value is obtained. Pooling all samples from Tunisia, Turkey and Sudan or treating the Tunisian and Turkish samples together and measuring I^S_A gave low positive values (0.0120 and 0.0095, respectively) suggesting a degree of linkage disequilibrium (Table 2.8.). The null hypothesis of linkage equilibrium was tested by calculating V_D and L, as described in Section 2.2.4. V_D was calculated to be greater than both values for L (L_{MC} and L_{PARA}) supporting the conclusion of linkage disequilibrium (LD) over these combined populations. While there are a number of reasons why LD could be observed, the simplest hypothesis was tested – i.e. that these populations are geographically sub-structured. The values of I^S_A obtained for the separate populations from Turkey and Tunisia (Table 2.8.) are close to zero with $L > V_D$, suggesting that each population is panmictic (i.e. undergoes random mating) and supporting the hypothesis that there is geographical sub-structuring. The loci analysed were either on separate chromosomes or separated by large physical distances on the same chromosome (Figure 2.4.) except for markers TS8 and TS9, which are 38 kb apart. To confirm these loci were genetically unlinked, the I^S_A of the Tunisian population (n = 51) was re-calculated using only these two markers and a value of -0.0154 was obtained, indicating no discernable association between alleles at these loci ($L > V_D$ supporting linkage equilibrium). This data supports a high level of recombination consistent with the unit of recombination being of small physical size. The limited number of stocks from the other countries represented population samples of insufficient size for meaningful comparison and therefore these stocks were not analysed using this method.

2.3.5. Population sub-structuring and diversity

A high level of genetic diversity was observed in stocks isolated from across Tunisia and Turkey and also across regions within each country. This was indicated by the mean estimated heterozygosity (H_e), which is calculated from the allele frequencies within each group and is presented in Table 2.8. This ranged from 0.913 and 0.929 in the two Turkish provinces to between 0.858 and 0.867 in the regions within Tunisia. This indicated that the Turkish population was slightly more diverse, although somewhat paradoxically they exhibited a lower mean number of genotypes per stock than the Tunisian population. This anomaly was attributed to H_e and 'mean number of genotypes per sample' measuring different 'types' of diversity; the first was across the population while the second was

within the individual host. In part, the finding may be related to the fact that several of the Turkish cell lines had been subject to more prolonged passage than Tunisian stocks.

To test the conclusion that the populations in each country were genetically separate but that there was limited sub-structuring within the populations, the genetic differentiation between populations was assessed by measuring the reduction in heterozygosity, i.e. F_{ST} values. Two estimators of F_{ST} (G_{ST}' and θ) (Weir and Cockerham 1984; Nei 1987) were calculated, producing consistent results for all combinations of populations analysed and are also presented in Table 2.8. A moderate amount of differentiation ($G_{ST}' = 0.056$, $\theta = 0.058$) was evident among the three countries and similarly a moderate level of differentiation was indicated when the populations from Tunisia and Turkey were analysed ($G_{ST}' = 0.044$, $\theta = 0.045$). The power of a multilocus genotyping system to differentiate between populations lies in the number of markers and the allelic distribution of each marker. To assess the ability of each locus to differentiate between Tunisian and Turkish populations, the value of G_{ST}' and other indices of diversity were evaluated for each marker and are shown in Table 2.9. G_{ST}' values of around zero indicated a lack of differentiation and were obtained for TS5, TS15 and TS31 while the largest value of 0.188 was demonstrated by TS25, indicating moderate differentiation. Across the set of markers, a relationship is clearly evident between the level of gene diversity (i.e. estimated heterozygosity) and the ability of that marker to discriminate between populations. That is to say, markers with lower diversity within a population show most evidence of differentiation between populations. This correlation is plotted in Figure 2.8. ($r^2 = 0.51$, $p < 0.05$). In contrast with the differentiation evident between different countries, when the populations from single countries were analysed independently, F_{ST} values dropped significantly, giving a low level of differentiation within Tunisia, while the Turkish samples gave a negative value for F_{ST} (Table 2.8.). Although F_{ST} values should not be negative, if the 'true' value is close to zero, the estimated value from a small sample will vary around that value and hence sometimes will be negative. Consequently, negative values can be interpreted as being zero and indicating no genetic differentiation.

These results are consistent with the population analysis reported in the previous section. F_{ST} values of 0.009 were obtained for the three largest sampling sites within Eastern Tunisia (data not shown), indicating a further drop in differentiation when examining populations within one region of Tunisia compared to the country as a whole. This supports the conclusion that the parasites in Tunisia may be regarded as a single panmictic population.

Table 2.9. Indices of marker diversity and differentiation

The degree of differentiation between the populations of *T. annulata* in Tunisia and Turkey was determined by comparing within-sample gene diversity (H_S) with overall gene diversity (H_T'). Averaged over all ten loci, F_{ST} was estimated at 0.044, implying a moderate amount of genetic differentiation between populations.

Table 2.9. Indices of marker diversity and differentiation

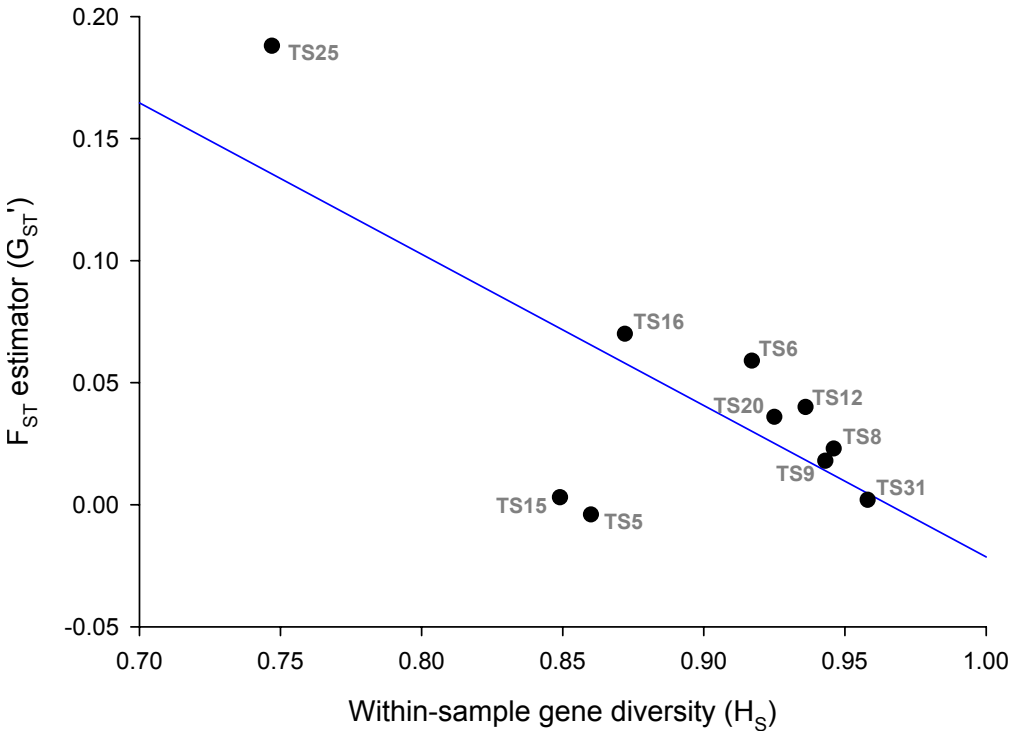
Locus	H_S	H_T	H_T'	G_{ST}	G_{ST}'
TS5	0.860	0.859	0.857	-0.002	-0.004
TS6	0.917	0.946	0.974	0.030	0.059
TS8	0.946	0.957	0.968	0.012	0.023
TS9	0.943	0.952	0.961	0.009	0.018
TS12	0.936	0.956	0.976	0.021	0.040
TS15	0.849	0.851	0.852	0.002	0.003
TS16	0.872	0.905	0.938	0.036	0.070
TS20	0.925	0.942	0.959	0.018	0.036
TS25	0.747	0.834	0.920	0.104	0.188
TS31	0.958	0.959	0.959	0.001	0.002
Mean	0.895	0.916	0.936	0.022	0.044

H_S = within-sample gene diversity, H_T = overall gene diversity,
H_T' = overall gene diversity (independent of number of samples), G_{ST} = estimator of F_{ST}
G_{ST}' = estimator of F_{ST} (independent of number of samples)

Figure 2.8. Regression analysis of within-sample gene diversity against an estimator of F_{ST}

Estimated heterozygosity, or within-sample gene diversity (H_S), was correlated with a measurement of population differentiation, G_{ST}', between Tunisian and Turkish samples. The 'least-squares' line, as derived from linear regression analysis is shown in blue, demonstrating an inverse correlation between these two parameters.

Figure 2.8. Regression analysis of within-sample gene diversity against an estimator of F_{ST}



$r^2 = 0.51$, $p < 0.05$

To quantify the effect of geographical separation on population differentiation, Nei's genetic distance (D) (Nei 1978) was calculated between populations from Eastern, Western and Central Tunisia, Northern and South-western Turkey and Sudan. Values ranged from 0.28 between Eastern and Western Tunisia to 2.04 between the Sudanese and Tunisian populations. Regression analysis (Figure 2.9.) was undertaken to investigate the relationship between genetic and geographical distance between populations and a positive correlation observed ($r^2 = 0.809$, $p < 0.0001$). A similar positive correlation was demonstrated by regression analysis of pair-wise F_{ST} values against geographical distance ($r^2 = 0.52$, $p = 0.002$, data not shown). An additional, novel method was used to demonstrate the relationship between genetic differentiation and geographical separation of populations. Triangular similarity matrices were constructed for the data representing (i) the mean geographical distance (km) between the sampling regions and (ii) the mean value of D between regional populations. These were used to construct a pair of dendrograms, which are presented in Figure 2.10. These display almost identical topology supporting the hypothesis that genetic differentiation correlates with physical distance separating populations.

Unfortunately, these genetic measurements are based on relatively small samples of *T. annulata* populations, especially in Turkey. Consequently, the results must be interpreted with a degree of caution, and should be viewed as preliminary evidence for differentiation of geographically separated populations. In order to use these particular standard genetic tools to investigate this issue, further studies are indicated using a much larger number of samples collected specifically for this purpose. However, in the meantime, similarity analysis allows an opportunity to analyse this limited data set in order to support the hypothesis of geographical sub-structuring of *T. annulata* populations.

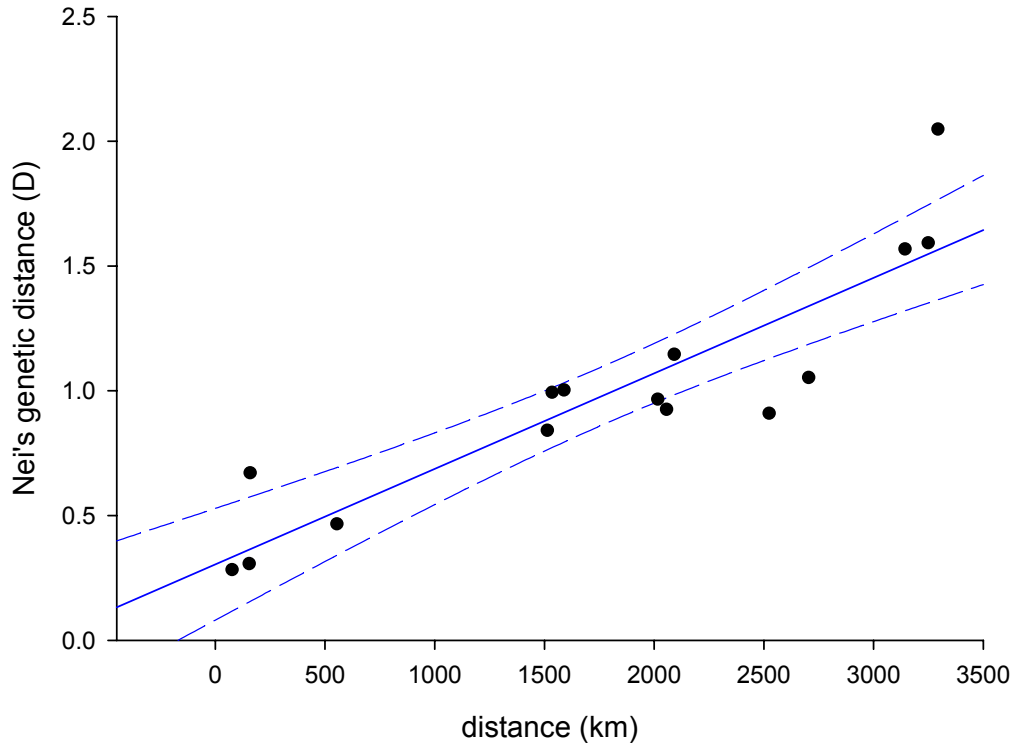
2.3.6. Similarity analysis of geographically separate populations

Due to the high level of diversity within *T. annulata* and the large number of mixtures of genotypes present in individual stocks, it was necessary to construct a 'genetic fingerprint' encompassing the entire allelic profile of each sample in order to demonstrate genetic relationships between parasite populations from different geographical areas. One hypothesis is that the population structure is panmictic and that sexual recombination occurs at a sufficiently high rate, such that discrete genotypes are not stable over time. The mixture of genotypes ingested by a tick feeding on the cattle host represents a sub-population, which has the ability to sexually recombine in the tick gut. Hence, this mixture of genotypes was treated as an operational unit for making comparisons between samples.

Figure 2.9. Nei's genetic distance as a function of geographical separation of populations

Nei's genetic distance (D) (Nei 1978) was calculated between populations from eastern, western and central Tunisia, northern and south-western Turkey and Sudan. These values were compared to the geographical distance (km) between sampling areas and a statistically significant positive correlation was demonstrated using linear regression analysis. The 'least-squares' curve is represented by a solid blue line and the 95 % confidence interval is represented by the two dotted blue lines.

Figure 2.9. Nei's genetic distance as a function of geographical separation of populations



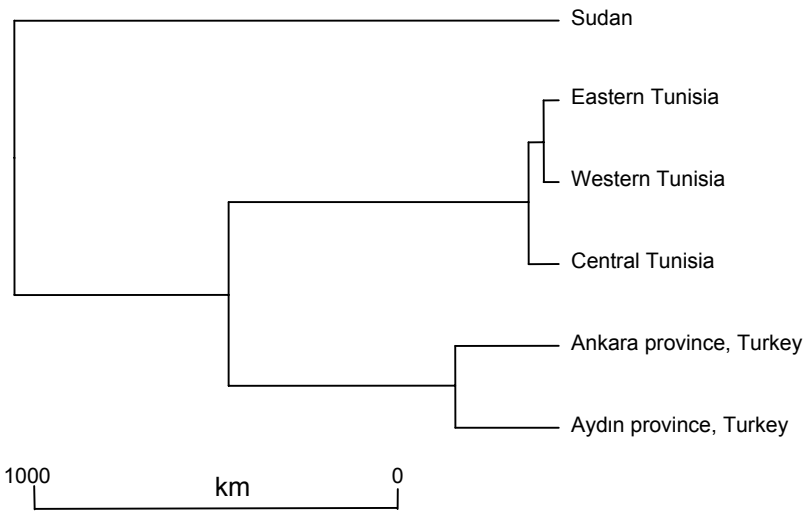
$r^2 = 0.809$, $p < 0.0001$

Figure 2.10. Dendrograms of distances between populations

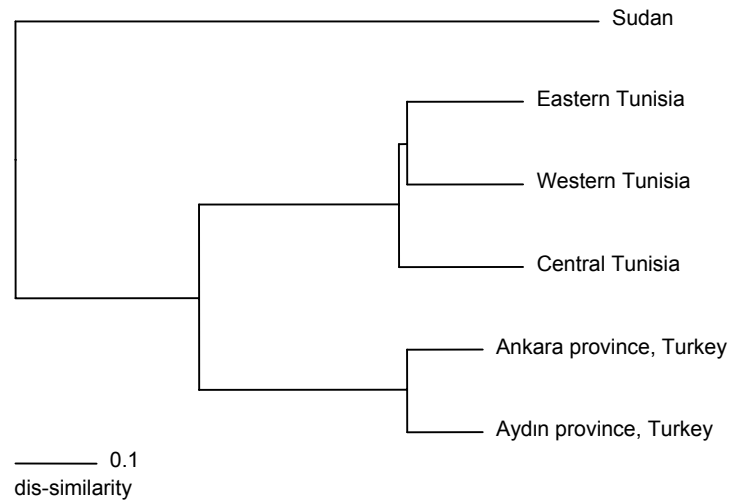
Lower triangular Euclidean similarity matrices were constructed for data representing **(i)** the mean geographical distance (km) between the sampling areas and **(ii)** the mean value of Nei's Genetic distance (D) between regional populations in Tunisia, Turkey and Sudan. In both cases, clustering of entries was performed using an unweighted arithmetic average method and the results used to construct a pair of dendrograms, which displayed almost identical topology. Together with Figure 2.9., these data support the hypothesis that genetic differentiation correlates with physical distance separating populations.

Figure 2.10. Dendrograms of distances between populations

(i) Geographical



(ii) Nei's genetic distance (D)



Comparison of principal MLGs alone did not provide conclusive evidence of geographical associations between samples, as the data set was not 'deep' enough (data not shown). That is to say, the degree of marker polymorphism was so great that few alleles attained a high frequency. Less polymorphic markers or a greatly increased sample size would have improved the comparison of MLGs. Therefore, Jaccard's co-efficient was used to determine the similarity between the total allelic profiles of each sample. This information was used to construct a dendrogram, using an unweighted arithmetic average method (Figure 2.11.). The general structure of the tree is clear: samples from individual countries clustered together, not withstanding a small number of outlying samples, which confounded the geographical clustering. The stability of this tree was tested by bootstrap analysis using 1,000 pseudo-replications. The node separating Sudanese and three Turkish stocks from the rest of the samples, gives a value of 1,000. Towards the extremities, values above 900 were obtained, where samples from the same country clustered together. However, there was less stability towards the root of the tree. This was attributed to samples from Eastern, Western and Central Tunisia being interspersed. The same is seen with Turkish samples from different areas. In other words, the dendrogram does not resolve subtle differences between samples from within the same country. Therefore, only macro-geographical sub-structuring can be inferred from this analysis. To confirm the ability of the analysis to cluster putatively geographically closely related samples, the tree was expanded to include piroplasm extracts and cell line cultures derived from the same animal (data not shown). As predicted, such pairs of samples clustered together.

This analysis provided clear evidence of geographical sub-structuring within the *T. annulata* population, which was previously quantified when Nei's genetic distance and Wright's fixation index values were correlated with geographical distance between sampling sites.

2.4. Discussion

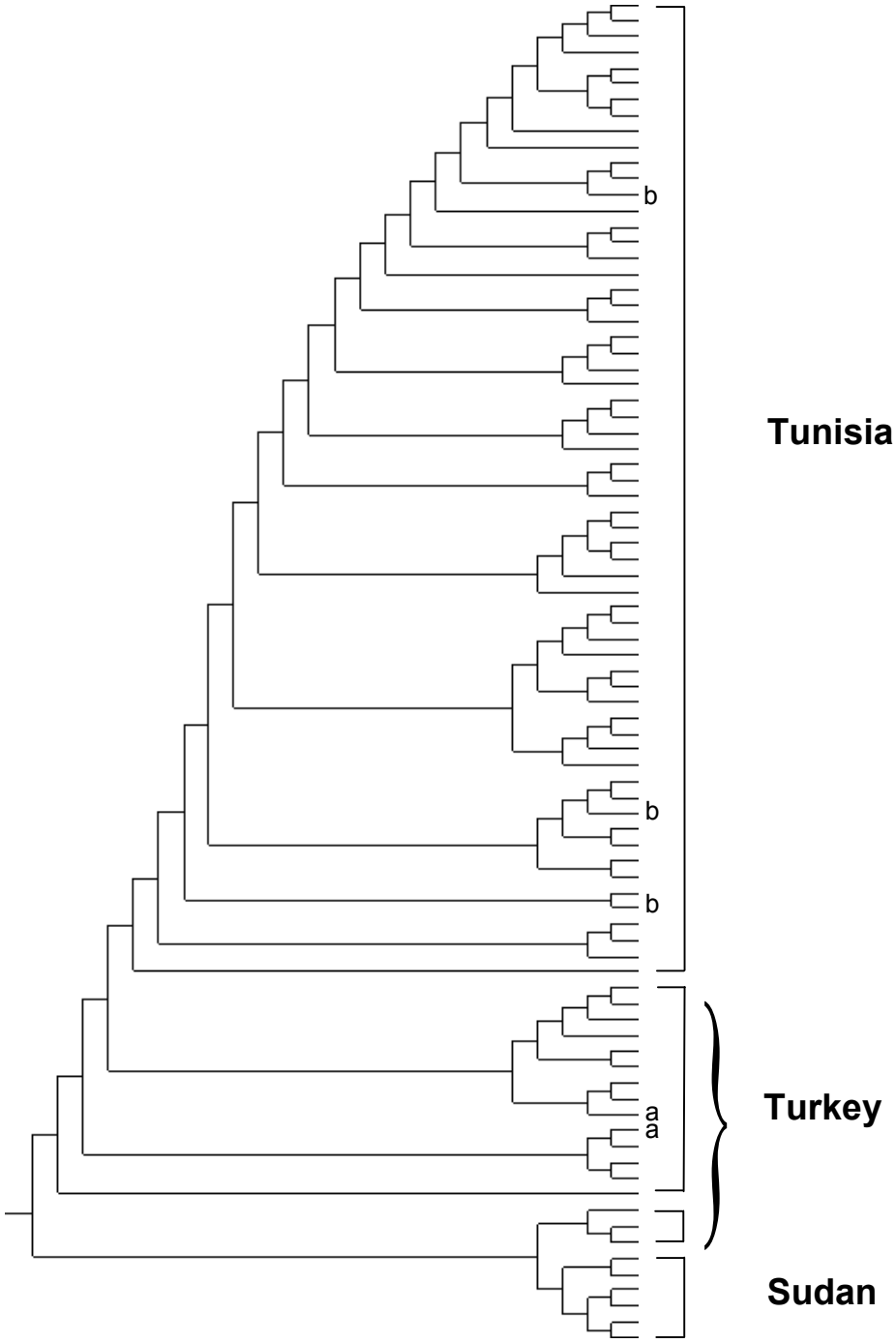
2.4.1. Suitability of markers for genetic analysis of *T. annulata*

Marker based genetic studies are invaluable tools for investigating the basic biology of pathogens and the interplay between control strategies and the behaviour of the organism at a population level. Micro- and mini-satellite loci have proved to be informative markers for such population studies, but they are not without their limitations. Polymorphisms in the primer annealing sites may lead to 'null alleles', a consequence of failure to amplify PCR product. In this study, six markers were able to amplify from every single parasite

Figure 2.11. Jaccard's similarity of allele fingerprints

Jaccard's co-efficient (Jaccard, 1908) was used to determine the similarity between samples using their complete allelic profiles. This data was clustered by an unweighted arithmetic average method and used to construct a dendrogram. The dendrogram is presented as a rectangular cladogram solely to illustrate the tree topology and therefore a scale is not indicated. Macro-geographical sub-structuring could be inferred, however bootstrap analysis demonstrated that the tree was relatively unstable except towards the extremities, where isolates from the same country clustered (values not shown).

Figure 2.11. Jaccard's similarity of allele fingerprints



a = Tunisian outliers

b = Turkish outliers

stock and an overall failure rate of less than 0.7 % of reactions was observed across all markers. This very low rate is attributed to the efficiency of the screening method and ensures a virtually unimpaired genetic dataset. Markers suitable for population studies should ideally be neutral and genetically unlinked. Large physical distances separate all markers except TS8 and TS9 (Figure 2.4.), although it has been demonstrated that even these loci, though only 38 kb apart, are genetically unlinked. This suggested a high recombination rate, which is consistent with that calculated from genetic crosses in *P. falciparum* where 1 cM = 15-30 kb (Walker-Jonah *et al.* 1992).

The published genome sequence indicated that nine markers are beside or within predicted proteins (four hypothetical, five conserved hypothetical). The results suggest the pattern of divergence may be constrained within some exon-located repeats (TS5, TS8 and TS24), however it does not follow that such regions are under selection. Selection acting on a particular marker would be predicted when the locus was in or near sequence representing an allele that is more or less advantageous to the organism. These three markers displayed polymorphism, which was consistent with the maintenance of an open reading frame (ORF). Insertions and deletions that disrupt the ORF may result in non-viable parasite genotypes, and in a sense such deleterious mutations would be selected against. However, this may be viewed as a form of constraint on the polymorphic capability of the locus across the whole parasite population. This is quite different from the selective pressures associated with phenomena such as drug resistance or immune evasion. For example, if particular marker were in proximity to an antigen gene under a selective pressure, the marker sequence would also be under selection whether it was in the coding sequence or in non-coding sequence, because it is genetically linked to the locus under direct selection. In general, therefore, whether a marker lies within coding or non-coding sequence does not indicate whether it is under selection or not.

Unquantifiable stochastic effects may influence allele frequencies for any marker and the resulting measures of differentiation, therefore SE values for F_{ST} estimator θ were included in Table 2.8. Values for G_{ST}' for each locus were calculated and found to be normally distributed among all markers over all population combinations analysed (Kolmogorov-Smirnov normality test). Furthermore, a relationship was demonstrated between the amount of diversity displayed within a population and the ability of that marker to detect differentiation between populations (Table 2.9. and Figure 2.8.). For example, marker TS25 showed the highest level of differentiation by far (F_{ST} = 0.188) while it showed the least within-sample divergence. This was mirrored in the allele frequency data

(Figure 2.7., TS25 locus) where a low number of different alleles were present at a relatively high frequency in each population. Such markers, with limited numbers of alleles present at high frequency are considered ideal for population genetic studies as they will provide a more stringent test for linkage equilibrium in populations with a small sample size.

The complex mutation pattern associated with several of these markers is analogous to the situation described in a study of *P. falciparum* (Anderson *et al.* 2000b). This analysis showed that many micro-satellite loci deviate from the conventional step-wise mutation model, precluding the use of analytical tools based on this mechanism. Since the primary aim of this study was to investigate the genetic structure of current *T. annulata* populations without drawing any phylogenetic inferences, definitively fitting the data to a particular mutation model was unnecessary.

2.4.2. Population structure

A large number of genotypes was demonstrated in the majority of samples. This reflects the situation in *T. parva*, where mixtures of genotypes infecting individuals are very common (Oura *et al.* 2003). The data suggests that ticks feeding on a host are likely to ingest a mixture of genotypes, which have the capacity to sexually recombine within the vector. In *Plasmodium*, the amount of recombination has been related to transmission intensity, with areas of low transmission showing evidence of self-fertilisation and inbreeding, whereas in areas where transmission intensity is high, increased levels of outbreeding are demonstrated (Anderson *et al.* 2000a). Consequently, in the geographical regions analysed in the present study, a panmictic structure may be predicted in *T. annulata*, as a high number of genotypes in each sample would imply high transmission intensity and provide the necessary background for such a mating system. In Tunisia, most farms are relatively isolated from each other and because of the life-cycle of the vector tick being postulated to be largely confined to the barns within each farm, it would be reasonable to predict a high level of geographical sub-structuring. Clearly the data presented here do not support this and suggest that there is movement of the parasites both between farms within a region and across regions. Given the tick has a limited range in the absence of its bovine host, cattle movement may be responsible for disseminating parasite genotypes throughout the country. An alternative hypothesis is that there is an extremely high level ancestral polymorphism maintained within separate *T. annulata* populations. That is to say, the multitude of extant alleles displayed by each marker may have been circulating in the general population for a very long period of time. If current genetically

isolated populations of *T. annulata* are sufficiently large, it would take considerable time before they display divergence, as the result of genetic drift. This would result in a high level of diversity being maintained in areas that are effectively genetically isolated. In future studies, it would be interesting to identify a number of farms that maintain 'closed herds', i.e. where the stock are maintained by interbreeding and there is no influx of cattle from outside the herd. By analysing *T. annulata* genotypes found within these putatively isolated cattle populations, it would be possible to test whether geographic sub-structuring of parasite populations can be detected in the absence of host movement or whether ancestral polymorphism masks this effect.

Despite evidence for significant levels of recombination, a degree of genetic isolation was identified between countries (Table 2.8.). As expected, this was reduced when focussing on populations from different regions within both Tunisia and Turkey. Indeed, within both these countries, linkage equilibrium combined with a lower I^S_A and limited differentiation further suggests a panmictic population structure in each country. Thus, the results indicate a gradient of genetic differentiation over physical distance, which is shown graphically in Figures 2.9. and 2.10. Although any one member of the population may have the capacity to breed with another, in the absence of trade, geographical barriers between regions (such as the Mediterranean Sea between Tunisia and Turkey) isolate these populations from one another, allowing divergence through genetic drift. These differences were detected at the regional and the country level when comparing MLGs. In addition, a dendrogram constructed using Jaccard's similarity and based on the full allelic profile of each sample was able to distinguish between parasites isolated from different countries (Figure 2.11.). However, when an allele distance matrix (Bowcock *et al.* 1994) was constructed between all samples, there was no correlation with physical distance between sampling sites (data not shown).

As micro- and mini-satellites are relatively fast evolving, values for Nei's genetic distance (D) within species, even in higher eukaryotes, are large. For example, when the genetic distance between herds of Rocky Mountain bighorn sheep (*Ovis canadensis canadensis*) were compared, D values ranged from 0.377 to 1.414 (Forbes *et al.* 1995). The value of D between individual sites within Tunisia ranged from 0.28 to 0.67 with the average value between Tunisia and Turkey being 0.98. These values are similar to D measured over genetically diverse populations of *T. parva* in Uganda (Oura *et al.* 2005). Co-ancestry coefficient measurements of population differentiation (θ) have been made between several geographically diverse populations of *P. falciparum* (Anderson *et al.* 2000a). The value

calculated for *T. annulata* between Tunisia and Turkey ($\theta = 0.045$) is intermediate between the low values for *T. parva* between the geographically clustered countries of Uganda, the Congo and Zimbabwe (0.003 - 0.012) where no genetic differentiation is detected, and the higher values between populations in different continents (0.100 - 0.184 between Central Africa and South-east Asia).

In previous studies, diversity of the major merozoite antigen gene *TaMS* could not be associated with geographical location (Katzer *et al.* 1998; Gubbels *et al.* 2000b), However, polymorphism in *TaMS* may be associated with evasion of the bovine immune system and possibly balancing selection, therefore it cannot be considered a neutral marker for population analysis. The mosaic pattern of diversity of this molecule, which was determined to be the result of a high level of recombination, is consistent with the predicted population structure of panmixia from the present study.

A high level of mixed infection was revealed in both cell line culture and piroplasm preparations derived from infected cattle. In one cell line stock, marker TS8 identified ten alleles with other loci also exhibiting polymorphism, indicating that at least ten distinct genotypes were present in the original isolate. The actual number may be much higher, since recombinant genotypes may also exist. Mixed infections were identified in the vast majority of samples, with single genotypes only encountered in laboratory derived clones or cell lines, which had been subjected to multiple passages *in vitro*. This suggests infected cattle normally harbour a multiplicity of genotypes and that single genotype infection is rare. This agrees with previous studies using these stocks, where a high level of heterogeneity was identified (Ben Miled 1993; Ben Miled *et al.* 1994) in individual stocks. In studies on *T. parva,* a high level of mixed infection was also identified in both cattle and buffalo (Oura *et al.* 2003; Oura *et al.* 2005).

Although for a number of years there has been some evidence of sexual reproduction, including a crossing experiment using cloned strains of *T. annulata* resulting in recombinant genotypes (Ben Miled 1993), this study represents the first formal genetic analysis of field populations of *T. annulata* and confirms the role of genetic exchange in the field. Whether selective pressures are responsible for the high level of sexual recombination and population diversity indicated by this study remains to be determined. Although panmixia was indicated among the Tunisian stocks, a more extensive sample size is required to draw firm conclusions about the population structure in Turkey. As described in Section 1.7.1., several states of endemicity have been reported in Tunisia (Darghouth *et al.* 1996b) and in general, the stocks genotyped in this study originated from

areas of endemic stability. That is to say, cattle herds are subject to a considerable level of challenge from a high number of ticks and disease is confined to younger animals, indicating a high transmission intensity of the parasite. This situation is analogous to the areas of high transmission intensity in *P. falciparum*, where the population structure has been shown to be panmictic (Anderson *et al.* 2000a). It is important to note that these findings do not exclude alternative population structures in areas of low transmission intensity, not encompassed by this study, where self-fertilisation may play a major role and general population diversity has been lost. Further studies in regions of high endemic instability of tropical theileriosis are required to investigate whether such alternate population structures can be detected that could have implications for development of control strategies.

2.4.3. Relationship of *T. annulata* to *T. lestoquardi*

Closely related to *T. annulata, T. lestoquardi* is the causal agent of malignant ovine theileriosis and like *T. annulata* it is transmitted by ixodid ticks of the genus *Hyalomma*. The distribution of these diseases partially overlaps, as *T. lestoquardi* has been identified in South-eastern Europe, Northern Africa, the Near and Middle East and India (Levine 1985). While *T. lestoquardi* can only infect small ruminants, *T. annulata* infection in sheep is capable of developing to the schizont stage, although progression to the piroplasm stage has not been demonstrated (Leemans *et al.* 1999). Due to morphological similarity, PCR based techniques have been developed to differentiate these species in small ruminants. A *T. lestoquardi* species-specific PCR (Kirvar *et al.* 1998) and RFLP analysis based on the 18S rRNA gene (Spitalska *et al.* 2004) can discriminate between species. Computer simulations have suggested that for micro-satellite markers, between 50 and 100 loci may be required for reliable phylogenetic reconstructions (Takezaki and Nei 1996), however 39 micro-satellite loci have been used to accurately reflect species phylogeny in *Drosophila melanogaster* when compared with previously accepted methods (Harr *et al.* 1998). However, a study based on ten micro-satellites was unable to differentiate between humans and three primate species (Bowcock *et al.* 1994). The single stock of *T. lestoquardi* analysed in this study has alleles outside the size range for *T. annulata* for six out of the ten markers, which immediately implies genetic differentiation between the species. This is intuitively likely since recombination of *T. lestoquardi* with *T. annulata* is unlikely to occur since the latter species has not been demonstrated to produce piroplasms in infected sheep. For recombination to occur in the vector, a tick would require to be co-infected with both species from the same blood meal to allow syngamy of gametes. The genotyping results of the *T. lestoquardi* stock showed that five markers have alleles that

are shorter than any allele described in *T. annulata,* one marker has a larger allele, three markers where the *T. lestoquardi* allele is represented in the *T. annulata* population and one marker that fails to amplify. A phylogenetic tree inferred from the 18S rRNA gene sequences (Schnittger *et al.* 2003) clustered these two parasites very closely, implying they have only recently diverged. The fact that 90 % of the markers used here can amplify alleles from a single *T. lestoquardi* stock, while none can amplify from the other four species tested is consistent with the relatedness of these two species. Why the alleles of the single *T. lestoquardi* stock are significantly smaller than any of those in the *T. annulata* population across half of the marker loci is unknown. However, one possible hypothesis is that there has been less opportunity for recombination within the *T. lestoquardi* population. Ancestral sequences from which current micro- and mini-satellite loci evolved are likely to consist of a single or limited number of repeat motifs. Over time, tandem replication increases the copy number of the motif in the general population until a point of equilibrium may be reached, such that all the alleles display variance around a larger mean allele size. The population drifts towards this distribution, which is constrained by an unknown, unidentified mechanism. Since the divergence of species, a reduced rate of mutation in *T. lestoquardi* could be explained by (1) a longer generation time due to parasite ecology, (2) lower multiplicity of infection, resulting in a lower recombination rate and (3) an alternate population structure, such as clonality where low diversity is observed across successive generations. The markers developed in this study could be used to investigate these hypotheses by a population genetic analysis of *T. lestoquardi*.

2.4.4. Implications for vaccination and control

The panmictic population structure of *T. annulata* has important implications for the parasite's capacity to circumvent control strategies such as drug usage and vaccination. The abundance of mixed populations and the high level of genetic recombination greatly facilitate the generation of novel genotypes. Drug resistance alleles may disseminate more quickly than would be expected in a predominantly clonal population. Unpublished reports of buparvaquone resistance in *T. annulata* populations (M. Darghouth, personal communication) could direct research towards the development of novel compounds to combat the disease. Should additional drugs be employed in the future, the high level of recombination coupled with the lack of clear geographical sub-structuring, suggests that multiple drug resistant strains could arise.

Sexual recombination is an efficient method of generating variation within a population and may have implications for cell line vaccination. A high rate of genetic exchange raises

the possibility of recombination occurring between vaccine stocks that generate piroplasms and naturally infecting parasites leading to generation of virulent strains. However, there is no reason to suspect that such novel genotypes would be any more pathogenic than strains existing in the field. Although immunity induced by cell line vaccination is not genotype-specific, the degree of protection has been shown to be higher against homologous rather than heterologous challenge (Gill *et al.* 1980; Hashemi-Fesharki 1988; Darghouth *et al.* 1996a). In this and previous studies (Sutherland *et al.* 1996; Darghouth *et al.* 1996a), attenuation has been associated with a reduction in the number of parasite genotypes in the cell line used for immunisation. Using the micro- and mini-satellite markers, two of the Tunisian vaccine-cell lines, 'Batan deu' (Batan2) and 'Jedeida quatre' (Jed4) have been shown to each contain only a single genotype. It may be expected that extensive vaccination with a single genotype may exert a selective pressure on field populations of the parasite. There is, however, no evidence that widespread and prolonged vaccination campaigns have resulted in the generation of vaccine-resistant strains in the field in any of the areas of the world where cell line vaccination has been undertaken. In other words, a single genotype is effective in stimulating immunity in the face of a diverse parasite population, which has immense potential to generate further variation. This suggests that the immune mechanisms associated with the conferred protection are relatively insensitive to genotypic variation. This contrasts, perhaps, with the situation in ECF, where a cocktail of *T. parva* genotypes is necessary to stimulate protective immunity and also may contrast with the natural acquired immune response to *T. annulata*. In the case of *T. annulata* it would be interesting to characterise a selection of vaccine lines from different areas of the world and test for (1) heterogeneity and (2) relatedness to the field population within the country of origin.

The high rate of genetic exchange may assist in identifying regions of the genome that are under positive selection, such as antigen genes responsible for eliciting protective immunity. Using population genetic and comparative genomic methods, antigen candidates may be identified and studied to gain understanding of the mechanisms that the host employs to protect against infection and that the parasite employs to evade immunity and allow transmission. Additionally, high levels of sexual recombination may allow identification of genes involved in drug resistance. For example, in *P. falciparum*, an extensive panel of micro-satellite markers has been used to characterise selective sweeps associated with chloroquine resistance (Wootton *et al.* 2002). Levels of genetic diversity were shown to vary substantially among different regions of the parasite genome, with linkage disequilibrium (LD) detected around the chloroquine-resistance associated gene,

pfcrt. Moreover, a significant loss of diversity was also encountered at the *dhfr* gene, which is associated with pyrimethamine resistance (Pearce *et al.* 2005). This was detected across a 70 kb region around the most highly resistant allele of this gene, and was attributed to the widespread use of pyrimethamine to treat malaria in Southeast Africa. Therefore, the ability to detect the presence of LD provides the basis for mapping genes involved in drug resistance in *P. falciparum* and also potentially in other species such as *T. annulata*, which undergo frequent sexual recombination.

2.4.5. Evidence of *in vitro* selection

The analysis of the population genetics of *T. annulata* presented in this chapter is primarily based on using cell lines established in culture from infected animals. This raises the question of whether this method of sample collection is representative of the parasite population in the cattle. A clear difference was demonstrated between the multiplicity of infection in blood samples and homologous cell line preparations. The parasite DNA from the blood samples, which primarily represented the piroplasm life-cycle stage, was shown to be more diverse across all ten loci, while the cell lines representative of the macroschizont population, albeit after culture *in vitro* were less diverse. Can this be taken as evidence for *in vitro* selection? To answer this question, the way in which the cell line preparations were generated must be considered. Loss of parasite genotypes can occur through several mechanisms, which may occur consecutively – (1) during host sampling, (2) when the cell line is established and (3) during *in vitro* passage. *In vivo*, piroplasms are derived from merozoites, which are generated from circulating infected monocytes as well as infected cells in the lymph nodes and other tissues (Ilhan *et al.* 1998). Assuming all infected cells are competent to differentiate, piroplasms may reflect the full spectrum of parasite diversity within the bovine host. Furthermore, since piroplasms are relatively long lived they may minimise transient fluctuations in the parasite population. In contrast, *in vitro* cell lines are derived from PBMs, which may only be a subset of the infected cells within the host. These cells represent parasite-infected leucocytes in the peripheral circulation and thus it is possible that not all parasite genotypes will be contained in this sample. However, this presupposes there is selection of particular genotypes in PBM or a difference in parasite population through time in PBM, neither of which have been demonstrated experimentally. As infected PBMs are introduced into culture, it is possible that certain genotypes are perhaps better able to establish *in vitro* and this may reflect their relative proportions in the PBM sample. Moreover, it is known that *in vitro* a proportion of cell lines will differentiate before establishing, killing the host cell and terminating the cell line (Shiels *et al.* 1998). Following this initial reduction in diversity, certain parasite

genotypes may be better able to adapt to life *in vitro*. Such genotypes may be selected through continual passage in culture, further reducing heterogeneity. Although a reduction in parasite diversity has been demonstrated in this study, further research is necessary to determine the relative influence of these three mechanisms in promoting selection. Such a study may involve genotyping PBMs extracted directly from infected blood and analysing a series DNA preparations representing different passage numbers of infected cell lines. An alternative explanation for the higher multiplicity of infection in piroplasm extracts compared to homologous cell line preparations is the relative concentration of parasite DNA in each type of sample. Unfortunately, this could not be determined from the available data and consequently this parameter was not standardised. This issue could be investigated in future by the serial dilution of each DNA preparation to determine the minimum concentration of each that could successfully PCR amplify.

When the matched piroplasm and cell line multilocus genotyping results were examined (Figure 2.6.), a general trend was observed. The amount of diversity in the piroplasm sample negatively correlated with the number of shared alleles between the preparations. In other words, the more genotypes a piroplasm sample contained, the more dissimilar were the predominant MLGs generated from each preparation. Furthermore, the predominant MLG for the Ankara A_2 cell line was identical to one of its derived clones (Table 2.6.), with the alleles for the second clone present at lower levels. However, the clones showed only 30 % and 40 % identity to the piroplasm MLG. At least six genotypes were present in the piroplasm extract (as indicated by TS31), although when recombinants were taken into account the potential numbers rise exponentially. It must be appreciated that the cloned lines are a sample of the cell line and the cell line is a sample of the parasite population represented in the piroplasm preparation. Therefore, it is entirely possible that numerous genotypes, including the most abundant ones in the piroplasm are not present in the derived cell line and clones. Consequently, by deriving a MLG from a cattle blood preparation, one is merely sampling a potentially large population present in the host.

In the face of immense diversity within isolates, it must be anticipated that a degree of inaccuracy will be introduced when predicting the predominant MLG. This may be due to preferential amplification of certain alleles, which may be caused by several mechanisms including short allele dominance and polymorphisms at the primer sites. This issue was not investigated in this particular study, however it is acknowledged that mis-association of alleles to create chimaeric MLGs may confound linkage analysis. For example, if two particular alleles at separate loci amplified preferentially and were over-represented in the

population, a bias may occur in favour of linkage disequilibrium. In contrast, if certain alleles amplified poorly then random secondary alleles may be represented in the MLG, artificially promoting LE. However, mis-association of alleles would not affect the estimators of population differentiation, because they analyse loci independently. Heterozygosity is estimated solely from allele frequency in a given population sample, since the parasite is effectively haploid.

2.4.6. Application of markers to field samples

The micro- and mini-satellite markers were able to differentiate between stocks from different countries, in contrast to previous studies where antigenic loci failed to do so (Katzer *et al.* 1998; Gubbels *et al.* 2000b; Schnittger *et al.* 2002). Each micro- and mini-satellite locus has been shown to be polymorphic, genetically unlinked and the genotyping results have been shown to possess a high degree of reproducibility. This was achieved principally using high quality DNA prepared from cell lines. However, some adaptations are necessary if this system is to be used in future epidemiological studies.

The current method for scoring alleles relies on assessing all the allelic data in the population and manually 'binning' alleles. In the case of markers TS5 and TS15 where the size difference between alleles is both large and consistent this presents little problem. However, in the case of markers such as TS20 (Figure 2.5.(ii)), with small irregular differences between allele lengths this method is more subjective. Therefore, if this genotyping system is to be used by other researchers, a robust method for classifying alleles must be developed and refined. Additionally, future field studies are likely to rely on genotyping from cattle blood samples, where high levels of parasite genotype heterogeneity are anticipated and it is unclear how this factor will impact on linkage analysis. Therefore, it would be advantageous to confirm the linkage disequilibrium and genetic differentiation between Tunisian and Turkish populations already shown by the cell line study using a new collection of direct blood preparations. Cluster analysis thus far has failed to separate samples from different countries using the principal MLG. The hypothesis that this can be achieved using a larger dataset may also be tested. Additionally, direct blood samples should provide a more realistic representation of the level of heterogeneity in the field compared with cell lines. If detailed records accompany a blood sample, it may be possible to relate both the multiplicity and distribution of genotypes present to the phenotype of the host. Factors such as sex and breed may well play an important role, as different classes of stock may be under different management regimes and be subject to differential tick challenge. It would be particularly interesting to

quantify the relationship between age of host and multiplicity of infection. In both Tunisia and Turkey, tropical theileriosis is largely regarded as being an endemically stable condition, where cattle are subject to a high rate of challenge throughout their life. As discussed in Section 1.7., in certain areas the disease exhibits endemic instability, which is related to a lower intensity of parasite transmission. In the case of endemic instability then one may expect to observe an increasing number of genotypes over the lifetime of the host, through a trickle of infection. In contrast, in endemic stability, a large number of genotypes would be predicted even in the youngest animals. A null hypothesis of no difference in multiplicity of infection between different ages of stock may be tested, with a departure indicating endemic instability.

2.4.7. Future questions

The findings presented in this chapter demonstrate the development and initial application of a panel of micro- and mini-satellite markers to characterise stocks and isolates of *T. annulata*. The markers were shown to be specific to *T. annulata* in cattle-derived isolates and to exhibit considerable length polymorphism. A high level of mixed infection was found and preliminary population genetic analysis revealed a very high level of diversity across the parasite population. A moderate amount of differentiation was detected between stocks representing populations in Tunisia and Turkey and evidence was generated to support the hypothesis that the parasite population is geographically sub-structured. In Tunisia, a panmictic population structure was indicated, however sample sizes from other areas were too small to draw any firm conclusions. In order to confirm these findings and to further investigate diversity in field populations a further study was undertaken. Using an extensive collection of parasite isolates from Tunisia and Turkey, specific questions are addressed in the following chapter, namely -

- Can this genotyping system be developed to accommodate a high throughput of samples?
- Can macro-geographical sub-structuring be identified using the principal MLG when a larger sample of blood preparations is analysed using such a system?
- Can geographical sub-structuring be identified on a finer scale within Turkey?
- How does vaccine cell line composition relate to field populations of the parasite in Tunisia and Turkey?
- Does multiplicity of infection correlate with the host phenotype?

CHAPTER THREE

THE POPULATION GENETICS OF TUNISIAN AND TURKISH ISOLATES

3.1. Introduction

3.1.1. Rationale for further study

The initial study presented in Chapter Two characterised a panel of polymorphic micro- and mini-satellite markers, which were used to analyse a small set of samples to give a broad insight into the genetic diversity of *T. annulata* both within and between populations. The parasite material available at that time was particularly suited to the development of the genotyping system, consisting of a considerable number of isolates from diverse geographical origins. DNA preparations representing the Tunisian population were derived from cell line stocks generated in a structured sampling programme, however the collection of Turkish samples was small and the data relating to their provenance was incomplete. As a consequence of this, the results from this initial population genetic analysis were considered preliminary. The results suggested that parasite populations are geographically sub-structured between Tunisia and Turkey and in the former country, the parasite exhibits a panmictic population structure. In order to test the validity of these conclusions, an extensive analysis on a properly structured set of samples obtained from each country was necessary. In addition, analysis of such a set of samples would allow a series of important questions to be addressed, namely –

1. Can macro-geographical sub-structuring be identified using the principal multi-locus genotype alone? Although sub-structuring of populations between Tunisia and Turkey was identified in Chapter Two, it was only detected when comparing full 'allelic fingerprints' from each sample (Figure 2.11.). When using multi-locus genotype data, representing solely the predominant allele at each locus, stocks isolated from each country could not be distinguished. This was attributed to the availability of a limited number of samples, particularly from Turkey. However, failure to separate the samples with respect to their country of origin may have been in part due to the method of illustration, i.e. using dendrograms. Although dendrogram construction is an effective method for representing data elements which cluster discretely, this hierarchical approach is not particularly well suited to presenting data of a continuous nature. Consequently, an

alternative method for presenting multi-variant data, principal component analysis (PCA), was selected for this study.

2. Can geographical sub-structuring be identified between districts in Turkey? The population sample used in the initial, preliminary analysis was too small to draw any conclusions about geographical sub-structuring and genetic differentiation among populations of *T. annulata* within Turkey. Additionally, it was impossible to investigate the underlying population structure of the parasite in this country and to perform comprehensive linkage analysis. Therefore, an important feature of the new collection of samples was the inclusion of a large number of field isolates gathered from several locations within one area of Turkey.

3. In what way does multiplicity of infection correlate with the host phenotype? A very high level of mixed infection was documented in the initial study. By counting the number of alleles present at each locus, the level of heterogeneity in the DNA preparations was estimated. Over the course of the initial study, variation was observed in the mean number of alleles detected across the ten loci amongst the parasite preparations analysed. With some of the variation attributed to the different 'forms' of the parasite, i.e. clones, cell lines and piroplasm extracts, the question was raised – can variation in host variables explain differing levels of heterogeneity among field isolates? For example, do vaccinated cattle show a reduction in the number of harboured genotypes in comparison to unvaccinated cattle from the same area? To answer this and other questions, it was necessary that the new collection of samples was accompanied by extensive information relating to the host from which each sample was isolated.

4. How does the composition of cell lines developed for vaccination relate to field populations of the parasite in Tunisia and Turkey? Several cell lines have been developed in each country for immunising cattle against tropical theileriosis and vaccination is currently practised in Turkey using one of these preparations. The impact of vaccination at a population level is currently unknown and it would be of particular interest to discover the influence of vaccination on *T. annulata* diversity in field populations. For example, it is unknown whether the parasite population among vaccinated cattle is different from the population within unvaccinated cattle in the same area. With vaccine cell lines attenuated in their ability to differentiate, it is doubtful whether immunising genotypes could progress to the merozoite stage of infection and thus infect erythrocytes. Therefore it may be hypothesised that immunising genotypes may not be detectable in whole blood preparations from vaccinated cattle, where it is infected erythrocytes which

contain the bulk of parasite DNA in the form of piroplasms. To investigate this issue, the frequency of the alleles represented in an immunising preparation will be measured in collections of vaccinated and unvaccinated field isolates, to provide the first population level data on the effects of a cell line immunisation programme.

However, before any of these questions could be answered it was necessary to obtain a suitable collection of parasite material.

3.1.2. Field samples

Conducting a major sampling programme to obtain parasite material was beyond the scope of this project. Fortunately, over the last decade there has been a systematic collection of blood samples from cattle infected during the disease season in endemic regions of both Tunisia and Turkey. This material was made available for this study by the kind permission of Professor M. Darghouth (ENMV Sidi Thabet, Tunisia) and Dr T. Karagenç (Adnan Menderes University, Turkey). These collections of samples may be considered to be highly representative of parasite populations in the field, since they were collected from both clinical cases and carrier animals, with parasite DNA prepared directly from each bovine blood sample. This contrasts with many of the parasite preparations used in the initial study where DNA was extracted from cell lines, which had been subjected to *in vitro* culture. It was demonstrated in the previous chapter that *in vitro* culture results in a loss of genotypes compared to the primary isolate. Therefore, by analysing DNA preparations made from whole blood, the principal genotype identified likely reflects the most abundant genotype in the host circulation, as opposed to the one which has grown most successfully in culture.

3.1.3. Technical considerations

The initial study using the micro- and mini-satellite markers demonstrated a high level of mixed infection in both cell line and piroplasm-derived parasite preparations. It was shown that piroplasm preparations from infected individuals may exhibit a mean of between four and five alleles at each locus, which is in broad agreement with an earlier study where up to four GPI isoenzyme alleles were identified in the same preparations (Ben Miled *et al.* 1994). In fact, several of the piroplasm extracts exhibited up to seven alleles at the more polymorphic micro- and mini-satellite loci. However, both purified piroplasm and cell line DNA preparations may underestimate the true level of heterogeneity within an individual host. As discussed in Section 2.4.5., cell lines probably represent a sub-population of the parasite within the animal as they are solely derived from

circulating PBMs and they may be liable to the effects of *in vitro* selection while being established. In contrast, purified piroplasm extracts may contain a better cross-section of the genotypes present in an infected animal. The infection of each erythrocyte occurs by invasion of a merozoite generated from the differentiation of macroschizonts in host leucocytes. The available evidence suggests that, despite the *in vitro* demonstration of piroplasm replication within the erythrocyte (Conrad *et al.* 1985), no inter-erythrocytic cycle occurs in *T. annulata*. In other words, parasitised erythrocytes will represent the population of merozoites derived directly from macroschizonts. With erythrocytes existing in the circulation for around only 100 days, piroplasm infection depends on sustained production of merozoites from immunologically privileged sites and consequently, may reflect those macroschizonts that have differentiated most recently. In *T. annulata*, the synchronicity of macroschizont differentiation with respect to parasite genotype is currently unknown. Even though it may not represent all genotypes carried within the host, since the piroplasm is an obligate stage of infection, it is representative of the genetic material transmitted to the next parasite generation. Parasite DNA extracted from whole blood is likely to predominantly contain infected erythrocytes, but also macroschizont infected leucocytes. Hence, a similar or increased level of diversity is anticipated in whole blood with that observed in purified piroplasm preparations. This however presents a technical challenge to effective micro- and mini-satellite genotyping, i.e. in the identification of the major allele at each locus and the accurate determination of its size. As it is expected that a multitude of alleles will be encountered at each locus when analysing field populations, a reliable, reproducible protocol is required to define alleles. In the initial study, all electrophoretograms (traces) were examined individually and peaks corresponding to fluorescently-labelled amplicons were identified manually. This was feasible because of the relatively limited number of samples (< 150) and also because a considerable number of DNA preparations represented clones, with only a single allele present at each locus. Consequently, the traces exhibited only a moderate amount of complexity, displaying limited background 'noise' with a low PCR failure rate being attributed to the relative abundance of parasite DNA.

A current limitation of the genotyping protocol is that a separate PCR reaction is necessary to amplify from each locus in every DNA preparation. Adapting this protocol to a multiplex PCR system, whereby primer sets representing several markers are incorporated in a single reaction may seem an attractive prospect. A variety of fluorescent dyes are available for modification of oligonucleotides, and it would have been possible to label different PCR primer sets with different dye colours for co-amplification. However, this

was considered unfeasible due to the anticipated heterogeneity of the individual blood samples. Since most of the markers have large overlapping allelic size ranges and since a multitude of alleles at each locus is predicted, resultant traces would be very complex and difficult to interpret. Furthermore, differential sensitivities of the markers may result in preferential amplification of PCR products from more efficient sets of primers. This would result in a wide range of readings with amplicons in low abundance failing to be detected on the trace and resulting in a reduced identification of alleles by the less sensitive primers. In addition, overloading a Genescan™ assay mixture with respect to any one particular PCR product is likely to cause artefactual peaks to appear in alternative dye lanes further confounding the analysis.

Manually analysing and recording the data from a single Genescan™ trace takes on average five to ten minutes, depending on its complexity. This represents a considerable investment in time when one considers that to fully genotype 100 field samples across ten loci, manual allele identification would involve examining 1,000 Genescan™ traces. In *T. annulata* field populations with an average of five alleles per locus, this would involve recording the PCR product size and abundance (area under the curve) and would thus generate around 10,000 data points. Clearly, manually recording this volume of data is likely to result in a degree of transcription error. For these reasons, it was decided to adapt the genotyping system to allow automated allele identification by computer software. Additionally, the use of such software was likely to improve the reproducibility of genotyping results by removing extraneous information in a structured fashion. For example, certain PCR primer / template combinations are prone to generate an amplicon of predicted size and a second slightly larger amplicon, which incorporates an untemplated adenine (A) at the 3' terminus (Smith *et al.* 1995). Since, generally either the shorter or larger amplicon will predominate in such a reaction, the software may be programmed to consistently identify or 'call' this particular allele and ignore the other. Whether the larger or smaller allele is detected is immaterial, so long as the process is performed consistently. Commercially available genotyping packages are principally designed to analyse diploid data, and hence identify a maximum of two alleles at any locus. It was therefore necessary to select an application that could be programmed to harvest all the data from each of the ten loci from a large number of field isolates.

3.2. Materials and methods

3.2.1. Parasite material

3.2.1.1. Tunisian samples

87 bovine blood samples were collected between July 2000 and August 2003, representing clinical cases and carrier cattle primarily from two localities in Northern Tunisia - Béja and El Hessiène (see Figure 3.1.), a summary of which is presented in Table 3.1. Béja is located 100 km from Tunis in a fertile agricultural area between the Medjerdah River and the Mediterranean, beside the foothills of the Khroumire. All the samples from Béja were collected from clinically affected adult female cattle in August 2002, during the disease season. The other sampling site, El Hessiène is a village located in the Ariana region close to Sidi Thabet. Three neighbouring farms – Béchir, Hassine and Salah were sampled, representing mainly diseased Friesian–Holstein calves of between one and six months of age. In addition to the isolates from El Hessiène and Béja, a small number of samples (n = 16) were also obtained from Northern Tunisia. The exact origin of these isolates was unknown and for this reason they were included in only a subset of the population genetic analyses. These samples were isolated from cattle of a Friesian–Holstein type crossed with a local breed.

3.2.1.2. Turkish samples

Four districts in Aydın province in Western Turkey were sampled between 1996 and 2003, details of which are also presented in Table 3.1. The largest sampling site in the study was Sariköy village in Akçaova, which accounted for 52 of the 96 samples from this district, while 37, 30 and 38 samples were collected from Aydın, Incirlova and Nazilli districts respectively. The location of each of these districts is presented in Figure 3.1. The town of Aydın, lies in the centre of Aydın province, with the districts Akçaova, Incirlova and Nazilli located approximately 40 km south, east and west respectively. Both calves and adult cattle were sampled, some of which had been vaccinated with a cell line vaccine. These cattle consisted of mainly dairy type, Holstein and Brown Swiss along with a small number of indigenous breeds. A small number of additional samples were analysed (n = 17), including isolates from Kösk, Karpuzlu and Kuyucak. These fell outside the four main sampling districts and are denoted as *'Other Aydın province'* in Table 3.1.

Figure 3.1. New sampling sites in Tunisia and Turkey

(i) Tunisia

Parasite material was collected between July 2000 and August 2003 from two locations in northern Tunisia - Béja and El Hessiène. Samples from the village of El Hessiène represented the three neighbouring farms of Béchir, Hassine and Salah.

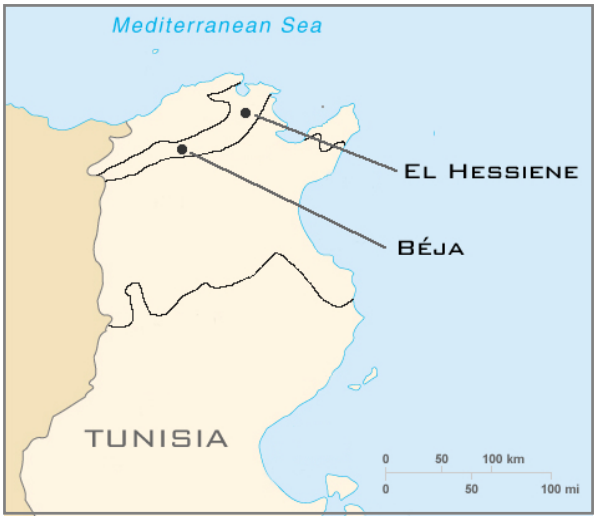
(ii) Turkey

Isolates were collected from the province of Aydın in the Aegean coastal region of western Turkey between 1996 and 2003 and represented the four districts of Akçaova, Aydın, Incirlova and Nazilli.

Details of the samples collected in each country are presented in Table 3.1.

Figure 3.1. New sampling sites in Tunisia and Turkey

(i) Tunisia



(ii) Turkey

Table 3.1. Parasite isolates from Tunisia and Turkey

A total of 305 parasite isolates were collected from northern Tunisia and western Turkey and the locations of these sampling areas are illustrated in Figure 3.1. The Tunisian samples were isolated principally from cattle in Béja and three neighbouring farms in the village of El Hessiène, representing mainly Friesian–Holsteins crossed with a local breed. *'Other (Northern Tunisia)'* refers to a limited number of samples where the precise site of isolation is unknown. The Turkish samples were obtained from four districts in the Aydın province on the Aegean coast of Turkey and were isolated from mainly Holstein and Brown Swiss cattle. Samples denoted as *'Other Aydın province'* were isolated outside these four main sampling districts; '*Other Akçaova'* and '*Other Nazilli'* refer to parasites isolated from within each district, but where the precise sampling site is unknown. Samples representing both '*Other (Northern Tunisia)'* and *'Other Aydın province'* were used to a limited extent in the genetic analysis. For each area in Tunisia and for each district in Turkey, the range, mean and standard deviation of the ages of the cattle are indicated together with their sex and vaccination status. Additionally, data corresponding to the Turkish villages represented by nine or more samples and the three farms in the village of El Hessiène are included.

Table 3.1. Parasite isolates from Tunisia and Turkey

Country	n	Area	n	Age (months)				Sex (n)			n vacc
				Min	Max	Mean	SD	M	F	ND	
Tunisia	**87**	**Béja**	**27**	**ND**	**ND**	**ND**	**ND**	**0**	**27**	**0**	**0**
		El Hessiène	**44**	**1.5**	**20.0**	**5.5**	**3.3**	**17**	**17**	**10**	**0**
		Béchir	*16*	*3.0*	*7.0*	*4.6*	*1.7*	*5*	*6*	*5*	*0*
		Hassine	*13*	*1.5*	*20.0*	*6.1*	*5.5*	*5*	*4*	*4*	*0*
		Salah	*15*	*2.0*	*12.0*	*5.8*	*2.3*	*7*	*7*	*1*	*0*
		Other (Northern Tunisia)	**16**	**1.0**	**5.0**	**3.2**	**1.0**	**9**	**7**	**0**	**0**
Turkey	**218**	**Akçaova district**	**96**	**6.0**	**180.0**	**29.3**	**26.2**	**54**	**39**	**3**	**32**
		Altinabat	*2*								
		Central	*2*								
		Kabalar köyü	*7*								
		Sariköy	*52*	*9.0*	*180.0*	*25.4*	*26.4*	*42*	*10*	*0*	*28*
		Other Akçaova	*33*								
		Aydın district	**37**	**8.0**	**120.0**	**37.8**	**25.7**	**14**	**23**	**0**	**1**
		Asagi Balikkoy	*6*								
		Balikkoy	*7*								
		Kadikoy	*1*								
		Kardes Koy	*2*								
		Kuyulu	*6*								
		Osmanbuku	*12*	*8.0*	*84.0*	*27.7*	*24.0*	*4*	*8*	*0*	*1*
		Sevketiye	*1*								
		Telsiztepe	*2*								
		Incirlova district	**30**	**3.0**	**129.0**	**60.0**	**34.7**	**2**	**24**	**4**	**1**
		Acarlar	*9*	*24.0*	*48.0*	*42.0*	*12.0*	*2*	*3*	*4*	*0*
		Hao	*21*	*3.0*	*129.0*	*63.4*	*36.6*	*0*	*21*	*0*	*1*
		Nazilli district	**38**	**3.0**	**138.0**	**49.8**	**35.9**	**6**	**32**	**0**	**6**
		Sümer Mah	*11*	*5.0*	*138.0*	*49.7*	*50.3*	*3*	*8*	*0*	*3*
		Güzelköy	*1*								
		Kestel	*10*	*3.0*	*81.0*	*38.4*	*25.1*	*0*	*10*	*0*	*2*
		Ocakli	*9*	*33.0*	*117.0*	*70.3*	*30.9*	*0*	*9*	*0*	*1*
		Other Nazilli	*7*								
		Other (Aydın province)	**17**	**12.0**	**96.0**	**30.9**	**29.8**	**2**	**4**	**11**	**1**

n = number of cattle sampled, Min = minimum, Max = maximum, SD = standard deviation,
M = male, F = female, ND = no data, vacc = vaccinated

3.2.1.3. Attenuated cell lines

High passage cell lines generated during vaccine development programmes in various countries were obtained for genotyping. This included the Tunisian cell lines Batan 2, Béja and Jedeida, which have never been deployed, and the Diyarbakir and Pendik cell lines, which have been used extensively in Turkey. A commercially available preparation, Teylovac™ (Vetal) was also included in the study, since this is the formulation commonly used in the Aydın region of Turkey. The datasheet for this product is presented in Figure 3.2. Vaccine cell lines were also obtained from India (Ode) and Spain (Caceres).

3.2.1.4. DNA preparation

EDTA blood samples taken from infected animals were frozen soon after collection and stored at -20 ºC. Between 100 µl and 300 µl of whole blood was thawed and the Wizard® Genomic DNA purification system (Promega) was used to prepare DNA according to the manufacturer's instructions. This involved chemically lysing red and white blood cells, after which cellular proteins were removed by salt precipitation to leave high molecular weight genomic DNA in solution. Genomic DNA was concentrated and desalted by isopropanol precipitation before being eluted in nuclease-free water and stored at -20 ºC. DNA from the attenuated cell lines was prepared as described in Section 2.2.1.

3.2.2. Automated genotyping

The ten micro- and mini-satellite markers developed and characterised in the previous chapter (Table 2.4.) were used to analyse each DNA preparation. PCR amplification was carried out using the conditions previously described in Section 2.2.3., including the incorporation of a fluorescently-labelled primer in the reaction. Amplicons were separated on an ABI 3100 Genetic Analyser in combination with the ROX-labelled GS500 standard size marker set. Genescan™ data files were then exported to proprietary software (Genotyper® 3.7., ABI) for the purpose of extracting genotypic information. For each of the ten markers, a custom software programme was developed in order to identify the peaks representing amplicons in each electrophoretogram. The following general method was used –

(i) Genotyper® software identified and 'labelled' up to twelve peaks within the reference range for each marker, i.e. from a lower limit of between 150 bp and 300 bp depending on the marker, to an upper limit of 500 bp.

Figure 3.2. Data sheet for Teylovac™

The data sheet provided with the commercial Turkish cell line vaccine, Teylovac™ is presented opposite. This vaccine is based on a high passage culture of the Pendik cell line, which was isolated near Ankara. Teylovac™ was commonly used to immunise cattle against tropical theileriosis in Aydın province, where the Turkish parasite material used in this study was isolated.

Figure 3.2. Data sheet for Teylovac™

VETAL

TEYLOVAC™

THEILERIA ANNULATA AŞISI

VETERİNER

KOMPOZİSYONU

TEYLOVAC, attenüe *Theileria annulata* (Ankara) suşundan doku kültüründe üretilmiş ve sıvı azot tankında taşınan bir canlı aşıdır.

ENDİKASYONU

Sığır, dana ve buzağılarda tropikal theileriosise karşı aktif koruma sağlamak için uygulanır.

KONTRAENDİKASYONU

Hasta olanlara ve ileri gebe hayvanlara yapılmaması tavsiye edilir.

AMBALAJ

TEYLOVAC aşısı:

a) 5 ml'lik cam şişelerde (flakon) 4-10 ve 15 doza eşdeğer lenfoid hücre içeren ve sıvı azot içinde dondurulmuş aşı

b) 15-50 ml'lik cam şişelerde 6.5 – 22.5 ve 35 ml sulandırma sıvısı ile uygulanmaya hazırlanır.

UYGULAMA VE DOZAJ

TEYLOVAC, sahaya sıvı azot termosları içinde dondurulmuş olarak taşınır. Uygulama için sıvı azot içinden çıkarılan 4-10 ve 15 dozluk aşı flakonu, önceden ıslatılmış pamuk kütlesi içerisinde veya kapalı bir kutuda 15-20 saniye bekletilerek, flakonu sızmalara bağlı olarak girebilecek, sıvı azotun neden olabileceği muhtemel patlamalardan korunulur. Akabinde 37 °C'daki su içinde çözdürülür. Çözdürülen aşı steril bir enjektörle kalın bir steril edinme iğnesi kullanılarak enjektöre iyice çekilip 6.5-22.5 ve 35 ml sulandırma sıvısı içeren şişeye boşaltılır. Yavaş şekilde enjektörle pompalanarak homojen bir dağılım sağlanır. Bu şekilde sulandırılmış aşı kullanıma hazır durumda 4-10 ve 15 doz aşı içerir.

UYARI

Tatbikata hazır sulandırılmış TEYLOVAC en geç **yarım saat** içinde kullanılmalıdır.

DOZ

Sulandırılmış aşının her bir 2.5 ml' si bir doz aşıya eşdeğerdir. Yaşlı hayvan, dana ve buzağılar için 2.5 ml'dir.

AŞI TATBİK YERİ

Asepsi ve antisepsiye uyulmak koşulu ile boyunda sağ ve sol preskapular deri altına tatbik edilir. Her hayvana tatbikattan önce şişede kalan sulandırılmış aşının enjektörle birkaç kez pompalanmak suretiyle homojen dağılımı sürdürülmelidir.

BAĞIŞIKLIĞIN OLUŞMASI (AKTİF KORUMA)

- Bağışıklık aşı uygulamasından ortalama 45 gün sonra başlar.
- Bağışıklık süresi en az bir yıldır.
- Aşı uygulanan tüm hayvanlara bir yıl ara ile tekrar aşı tatbik edilmek suretiyle bağışıklığın (aktif koruma) sürekli olması sağlanır ve enfeksiyon riski en alt düzeye indirilir.
- 3 Aylıktan küçük buzağılar hariç TEYLOVAC her ırk ve yaş gurubundaki hayvanlara tatbik edilir.

ÖNEMLİ UYARILAR

Ahır Theileriosis riskine karşı özellikle besi hayvanlarına da tatbik edilmelidir.

AŞI UYGULAMA ZAMANI

TEYLOVAC her iklim bölgesinin mevsimsel özellikleri ve tropikal theileriosisin epidemiolojisi dikkate alınarak en uygun zamanda yapılmalıdır. Özellikle taşıyıcı kenelerin *(Hyalomma spp)* mevsimsel etkinliğinin başlamasından en az iki ay önce o bölge hayvanlarına uygulanmalıdır.

Ege Bölgesi, Akdeniz Bölgesi, Güneydoğu Anadolu ile Iğdır ili çevresinde OCAK-ŞUBAT-MART ayları,

İç Anadolu, Marmara Bölgesi ve Karadeniz Bölgesinde ŞUBAT-MART-NİSAN ayları,

Doğu Anadolu Bölgesinde MART-NİSAN-MAYIS ayları TEYLOVAC tatbikatları için uygundur.

Her bölge için, uygulamalar önerilen aylardan önce veya 20-25 gün daha geç yapılabilir.

ZARARSIZLIK

TEYLOVAC tatbik edilen hayvanlarda hiçbir lokal ve genel yan etki oluşturmaz. Tamamen zararsızdır. Tatbikattan sonra görülecek enfeksiyonlar doğrudan hastalığın doğal inkübasyon periodu ile ilgilidir.

STABİLİTE VE MUHAFAZA

TEYLOVAC dondurucularda -70°C'de ve sıvı azot içinde -196°C'de en az 5 yıl aktif koruma gücünü yitirmeden muhafaza edilir. Sulandırma sıvıları (PBS) ise oda ısısında muhafaza edilir.

(ii) In order to eliminate minor amplification products, labels were removed from peaks whose height was less than an arbitrary value of 32 % of the maximum peak height. Removing a label meant that the peak was no longer analysed.

(iii) In order to suppress the '+A' effect, labels were removed from peaks which were preceded by a higher peak that was within 1.60 bp.

(iv) In order to reduce the effect of 'stutter' bands, labels were removed from peaks which were followed by a higher peak that was within 3.00 bp.

(v) For each sample, the position and area of every labelled peak was exported into a text file and subsequently imported into Microsoft Excel.

The predominant peak was defined as that with the highest area under the curve. For each locus using every sample, the predominant amplicon was ranked in order of ascending size (to two decimal places). That is to say, for each of the ten loci, a list of allele sizes encompassing the allelic spectrum was generated. Each of these ten lists was manually examined to allow the creation of fixed bins for the purpose of defining or 'calling' actual allele sizes. The performance and validation of this process is presented in the results section. Two types of data were generated from the analysis- (1) predominant allele multilocus genotype (MLG) data for each isolate and (2) multiplicity of infection data for each isolate, where the number of alleles at each of the ten loci was defined. Allelic size data for the non-predominant products of amplification was not included in the analysis. This was done to minimise the creation of artefactual alleles in the dataset and to allow a robust determination of allele frequencies. Consequently, calculations of allele frequency were based solely on the predominant allele at each locus.

3.2.3. Statistical analysis

The analysis of co-variance is often referred to by its acronym ANCOVA and is an alternative term for linear regression modelling using a single continuous explanatory variable along with one or more conditional factors. Analysis of co-variance of host variables, i.e. age (quantitative), sex, breed and vaccination status (all qualitative) was undertaken to determine which of these could explain the multiplicity of infection in individual cattle. This was achieved by comparing the average number of alleles present at each locus in an individual sample against the four host parameters using the statistical software package XLSTAT 2006.

3.2.4. Principal component analysis

Principal Component Analysis (PCA) is a powerful statistical technique for identifying patterns in multidimensional data, which has been applied to diverse fields such as facial recognition and digital image compression. PCA is used to reduce the number of dimensions in a dataset while retaining those characteristics of the dataset that contribute most to its variance. Essentially, a mathematical procedure transforms a number of potentially correlated variables into a reduced number of uncorrelated variables called *principal components*. The first principal component accounts for as much of the variability in the data as possible, with each successive component accounting for as much of the remaining variability as possible. The object of the analysis is to identify underlying trends within a dataset. The technique was applied to the multilocus genotyping (MLG) data generated in this study in order to identify patterns of distribution of parasite genotypes. The Microsoft Excel plug-in 'Genalex6' (Peakall and Smouse 2006) was used to construct a difference matrix and perform PCA on sets of MLG data. This software is freely available and can be downloaded at http://www.anu.edu.au/BoZo/GenAlEx/. The first two axes generated by each analysis was imported into and displayed by SigmaPlot8.0 (SPSS).

3.3. Results

3.3.1. Genotyping of field isolates

Genotyper® software was successfully used to genotype all 305 DNA preparations from the two populations with a full ten-locus MLG being generated for 257 (84 %) of these samples. Of the remainder, a single locus (generally TS9) failed to amplify in 34 cases, resulting in a nine-locus profile. The main features of allelic variation at each of the ten loci are presented in Table 3.2. Treating the Tunisian and Turkish populations as a whole, the mean number of alleles detected at each locus varied from 1.93 for marker TS16 to 3.54 for marker TS8. It can be observed that a maximum of twelve alleles were identified at any one locus, which was the 'cut-off' point for labelling peaks using the software. This value was reached in the case of TS6, TS8 and TS12, which represented three of the four most polymorphic loci, possessing between 49 and 61 defined alleles across the entire dataset. This association between the number of alleles detected per isolate per locus and marker polymorphism was intuitively correct, as highly variable loci should possess the greatest power to discriminate between the mixture of genotypes present in a single preparation. An ideal marker for population genetic analysis should amplify from all

Table 3.2. Allelic variation in the Tunisian and Turkish populations

All 305 new isolates were genotyped using the panel of ten markers. The minimum and maximum number of alleles detected at each locus within a single isolate was calculated across all ten markers for each isolate that amplified. The number of alleles represented in each population was calculated, taking into account only the most abundant alleles present in each isolate from that country. Gene diversity was calculated for each marker and is equivalent to estimated heterozygosity.

Table 3.2. Allelic variation in the Tunisian and Turkish populations

		TS5	TS6	TS8	TS9	TS12	TS15	TS16	TS20	TS25	TS31
Number of alleles within each sample	**Minimum**	1	1	1	1	1	1	1	1	1	1
	Maximum	8	12	12	8	12	9	6	9	9	10
	Mean	3.03	2.84	3.54	2.96	3.28	3.28	1.93	3.23	2.62	3.00
	No amplification	0	9	4	25	4	1	13	3	1	9
Number of alleles in population	**Tunisia**	8	28	28	25	33	8	20	21	6	23
	Turkey	12	50	48	32	44	11	30	28	17	54
	Overall	12	61	53	34	49	11	36	34	18	60
Gene diversity	**Tunisia**	0.827	0.946	0.946	0.960	0.956	0.829	0.898	0.885	0.708	0.886
	Turkey	0.818	0.950	0.963	0.947	0.963	0.808	0.858	0.872	0.671	0.967

samples. Markers TS9 and TS16 failed to generate a product for 8 % and 4 % of samples respectively, even after repeating the reaction using a reduced annealing temperature of 50 ºC. Across the other eight markers, a low failure rate of 1.3 % was observed. The most likely explanation for this low number of failures is sequence polymorphism at the primer annealing sites. In the case of TS16, primer-site polymorphism may also account for the low number of alleles detected at this locus (mean = 1.93). Where no amplicon was generated for a particular locus in an isolate, a null allele was entered into the multi-locus genotype. To investigate whether priming-site polymorphism was responsible for these failures, it would be necessary to redesign primers in more distant flanking sequence. The size of products amplified by these redesigned primers may be adjusted for compatibility with the results from the existing primer sets. Alternatively, amplicons could be sequenced and novel degenerate primers designed. However, the number of null alleles in the dataset was limited and the relatively low PCR failure rate did not warrant further investigation, given the objectives of the study. Generally, a high level of heterozygosity (gene diversity) was estimated for each marker in each population (Table 3.2.). The lowest value in each population was exhibited by marker TS25, with values of 0.708 and 0.671 for the Tunisian and Turkish populations respectively. Heterozygosity exhibited by the other nine markers ranged from 0.808 (TS15) to 0.967 (TS31) in the Turkish population.

To verify that the software was correctly processing the information generated during capillary electrophoresis, a number of traces were examined manually. Three example traces are presented in Figure 3.3., demonstrating the performance of the software to label peaks corresponding to PCR products generated by the marker TS20 using three different PCR templates. This marker served as a reasonable benchmark since a large number of alleles were identified in each sample, with a mean of 3.23 across the dataset (Table 3.2.). Figure 3.3.(i) illustrates the labelling of predominant peaks (in the blue trace), where the upper box beneath each peak indicates the size of the amplicon (in bp), while the lower box provides the area under the peak (in units^2). Two peaks were identified, while at least seven minor peaks were discounted. Whether these minor peaks represented genuine alleles is debatable – some peaks corresponded to known predominant alleles, while others were novel. Ultimately for the population genetic analysis this was irrelevant, since only the peaks with the greatest surface area were used to create MLGs. However, following use of the protocol detailed in Section 3.2.2., the number of alleles present at each locus in each sample was recorded and used to estimate the multiplicity of infection within each DNA preparation. The number of alleles identified at each locus therefore provided an index to the minimum number of genotypes present in the mixture. The ability of the

Figure 3.3. Example of automated genotyping (TS20)

Genotyper® software was programmed to identify and 'label' peaks in the electrophoretograms, which corresponded to fluorescently-labelled PCR products. The figure opposite illustrates the operation of the software to discard extraneous peaks from the analysis of three different isolates using the marker, TS20. The peaks which were removed from the analysis are denoted with an asterisk (*).

(i) Removal of low peaks

In order to eliminate minor amplification products, labels were removed from peaks whose height was less than an arbitrary value of 32 % of the maximum peak height.

(ii) Removal of '+A' peaks

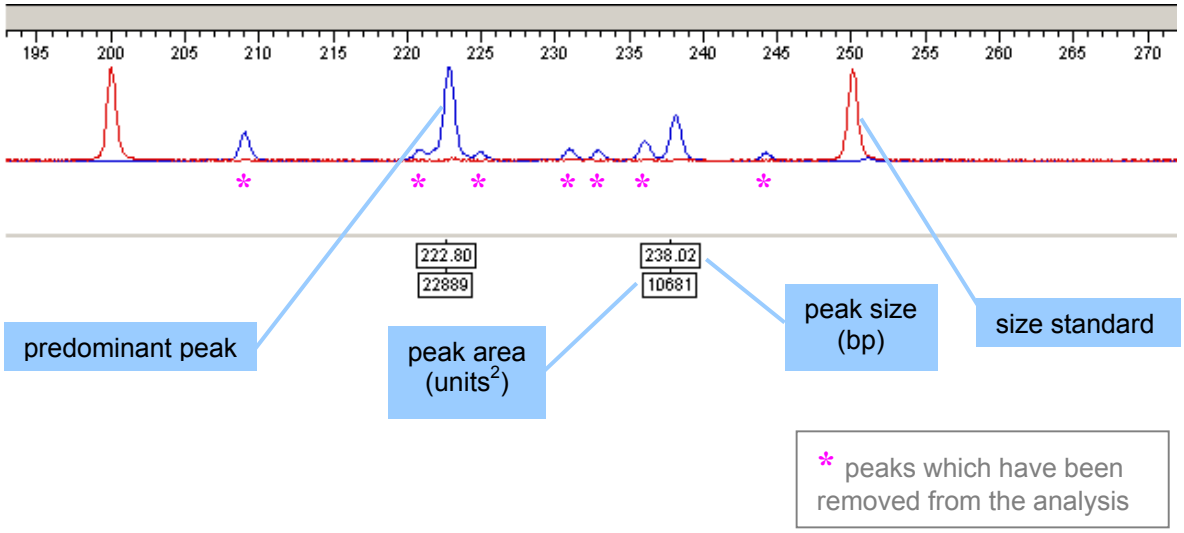
In order to suppress the effect of untemplated adenine nucleotides being added at the 3' terminal of PCR products, labels were removed from peaks that were preceded by a higher peak that was within 1.60 bp.

(iii) Removal of 'stutter' peaks

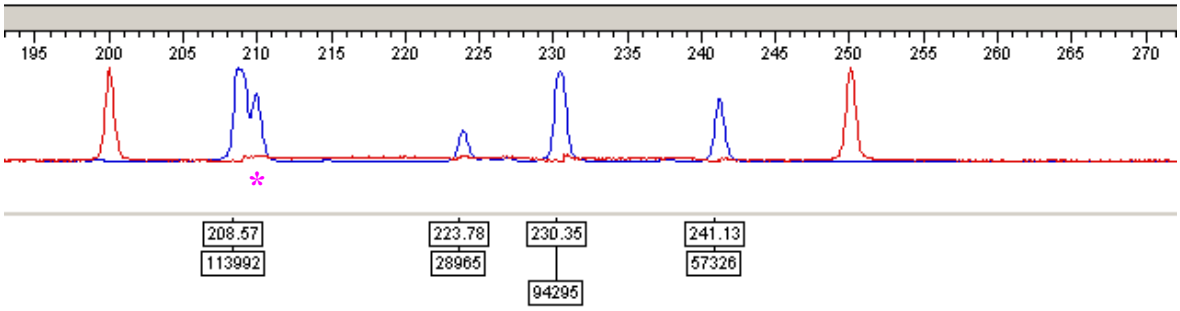
In order to reduce the effect of 'stutter' bands, labels were removed from peaks that were followed by a higher peak that was within 3.00 bp.

Figure 3.3. Example of automated genotyping (TS20)

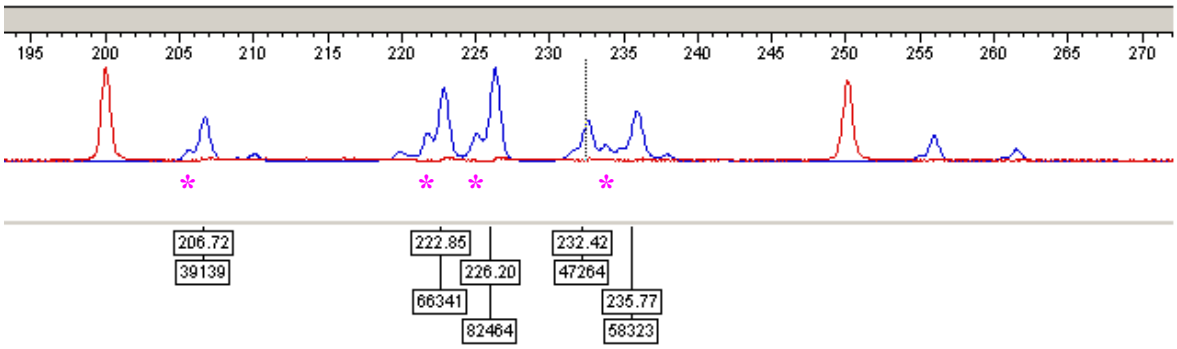
(i) Removal of low peaks



(ii) Removal of '+A' peaks



(iii) Removal of 'stutter' peaks



software to remove smaller peaks immediately preceding or following the major peak is also demonstrated in Figure 3.3. The removal of small peaks, significantly lower than the major peak is presented in trace (i), while an example of 'A+' peak removal is presented in trace (ii). Several peaks in trace (iii) show the presence of a small shoulder about one third of the height of the labelled peak. The first of the peaks may represent either a 'stutter' band or the 'true' PCR product, where the 'A+' effect was dominant. No matter which of these was the case, the important point was that the second of the peaks was labelled consistently. In general, there was a clear difference in the area of the major peak compared to secondary peaks and therefore the predominant allele could be easily identified.

Similar to the findings in the initial study, the markers showed differential patterns of polymorphism. In one case, a clear step-wise size difference between alleles was displayed, while in the other, the size variation was more continuous. Figure 3.4. depicts two markers at either end of this spectrum, TS5 and TS12. Marker TS5 consists of a perfect 6 bp repeat, as was demonstrated in the initial study (Figure 2.4.). Using the new larger dataset of predominant alleles, the allelic range for this marker was separated into 0.2 bp intervals, and the distribution of PCR product sizes was plotted (Figure 3.4.(i)). Clusters of PCR product can be seen at approximately 6 bp intervals, showing a normal distribution pattern around a central frequency value. For this marker, it was a straightforward task to define alleles since the bins were well separated, and a low variance was observed around a subjectively determined absolute allele size. Additionally, the results for this marker underlined the reproducibility of the genotyping system. The data presented in this figure represents several PCR assays, performed on different thermocyclers with Genescan™ analysis being undertaken over a period a several months. Due to the large sample size, several Genescan™ 'runs' were necessary for each marker. It can be concluded that the amount of variation caused through using separate PCR assays and sequencer 'runs' is low and that allele sizing is accurate to a level of < 1.0 bp. As can be seen in Figure 3.4.(ii)., the pattern of diversity of TS12 was quite distinct to that of TS5. 49 alleles have been identified across the dataset using this marker, ranging in size from less than 240 bp to greater than 360 bp. The distribution of PCR products is presented using 0.5 bp intervals. It can be seen that in the 250 – 280 bp range there was almost continuous gradation in allele sizes. The comparatively low number of alleles in each class and a lack of obvious clustering presented a challenge to accurate allele identification. Similar to the process used in the initial study, the raw predominant allele sizes were ranked in ascending size order and allele bins of around 1.0 bp were defined manually by

Figure 3.4. Allele binning

A clear stepwise size difference between alleles was displayed at several loci, while size variation was more continuous at others. An example of each is presented opposite.

(i) TS5 – perfect 6 bp repeat (0.2 bp intervals)

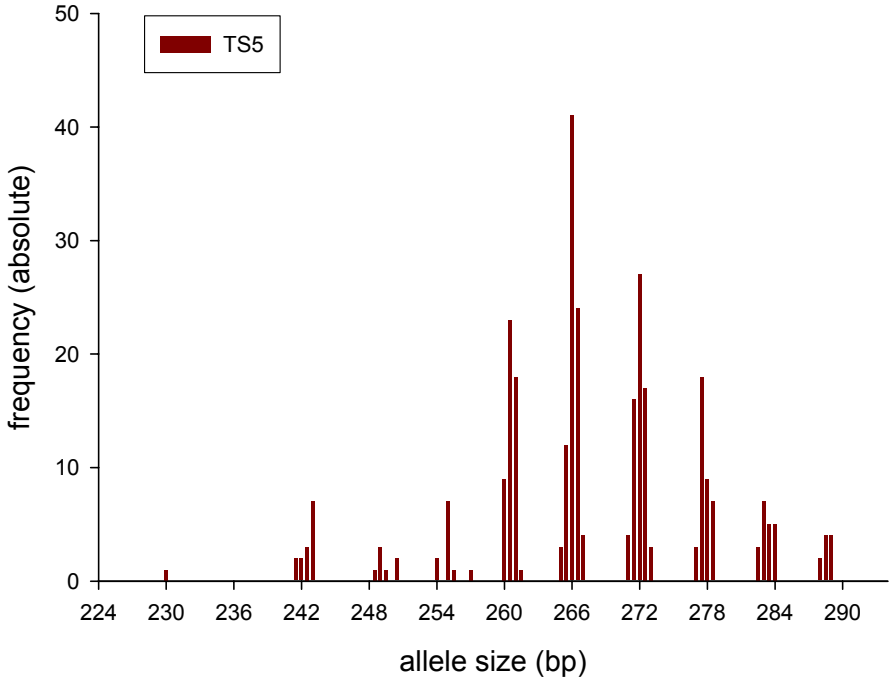
Alleles consistently differed in length by six base pairs, corresponding to the repeat motif, GGTTCA. An apparently normal distribution was observed in each PCR product cluster corresponding to each allele. Similar to the findings in the preliminary study, a limited number of alleles were present at a relatively high frequency in the population and the pattern of distribution remained consistent with a 'step-wise' mutation mechanism. Bins defining alleles at this locus were created at 6 bp intervals.

(ii) TS12 – variable repeat (0.5 bp intervals)

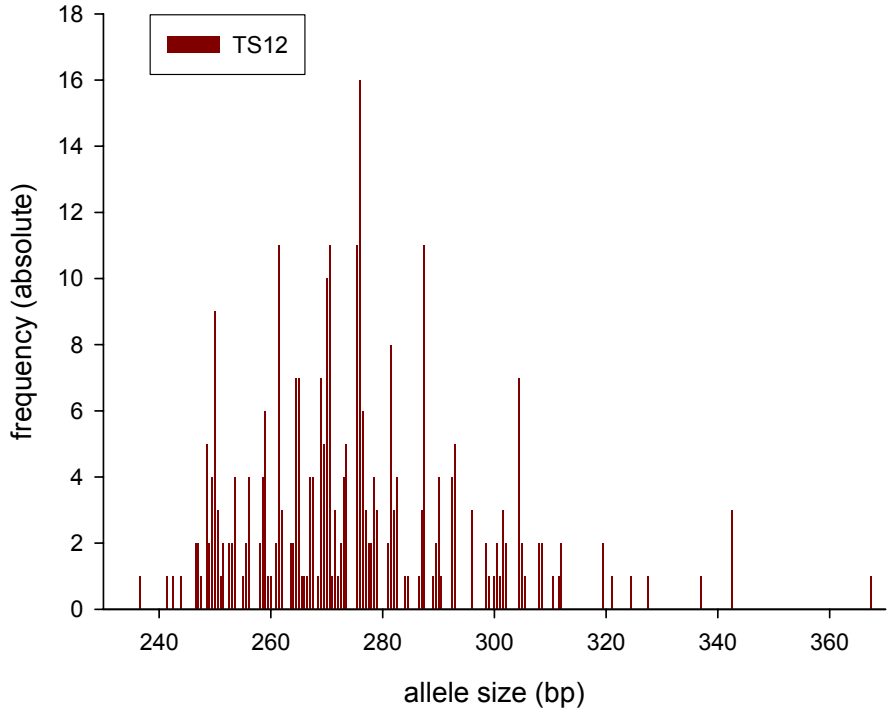
Alleles differed by a variable length, with a continuous gradation exhibited in parts of the range with a higher number of alleles observed at a relatively lower frequency in the population. This example illustrates a 6 bp micro-satellite motif, AATACT. The comparatively low number of alleles in each class and a lack of obvious clustering resulted in bins of approximately 1.0 bp being defined manually, following examination of the raw dataset.

Figure 3.4. Allele binning

(i) TS5 - perfect 6 bp repeat (0.2 bp intervals)



(ii) TS12 - variable repeat (0.5 bp intervals)



examining all the data points. Consequently for markers such as TS12, a large number of alleles were defined. The number of alleles detected for each marker in both Tunisia, Turkey and in the entire dataset is presented in Table 3.2. This ranged from 11 alleles for TS15 to 61 alleles for TS6.

For all ten markers, the frequency of each allele was determined in both countries and is presented in Figure 3.5. Similar to the initial study, a spectrum of allelic variation was observed across the markers when comparing the two populations. TS5 and TS15 again exhibited limited differentiation between populations, with similar allele frequencies in each country. In this study, four private alleles for the TS5 locus were found in the Turkish population with no alleles specific to Tunisia. This contrasted with the earlier analysis when a single private allele was detected in Tunisia (Figure 2.7.). This may be explained by the larger sample sizes used in this study, particularly that representing the Turkish population. The TS5 locus also demonstrated a general trend observed across the ten loci, i.e. additional alleles were identified when analysing this larger dataset. For TS5, the allelic range in the previous study was 246 - 288 bp, whereas in this study two additional alleles were identified at each end of the range, giving a new range of 228 - 342 bp. This was reflected in many of the other loci, where in addition to alleles outside the previously described range, many novel alleles were detected, intermediate in size between the alleles initially identified. This can be clearly observed when comparing locus TS6 (Figure 2.7. and Figure 3.5.). When the allele distribution for TS25 was compared with the earlier study, a common trend was observed, i.e. the 213 bp allele predominated in the Tunisian population (frequency ≈ 45 %), while the 218 bp allele predominated in the Turkish population (frequency ≈ 55 %) (Figure 3.5.). The calculation of allele frequencies at each locus enabled the level of gene diversity (estimated heterozygosity) to be calculated for each population. Interestingly, the level of heterozygosity for a particular marker was independent of the number of alleles it possessed. The estimated heterozygosities in both studies agreed with each other to a large degree and this is presented graphically in Figure 3.6.(i). That is to say, a range of estimated heterozygosities were observed across the ten loci and where a locus exhibited high or low heterozygosity in the initial study, this was reflected in this second, more extensive study. In the second study, it was also found that levels of heterozygosity between the Tunisian and Turkish populations were similar across each marker and this is presented graphically Figure 3.6.(ii). Therefore, in general it can be concluded that the estimated heterozygosity value for a particular locus is a consistent feature of that locus, and is relatively independent of sampling location. The unusually low estimated

Figure 3.5. New Tunisian and Turkish allele frequencies

The frequency of each allele in the new Tunisian and Turkish population samples was determined as a percentage (%) of the total for all ten markers. This data was solely based on the predominant allele in the 87 Tunisian and 218 Turkish isolates described in Table 3.1.

Figure 3.5. New Tunisian and Turkish allele frequencies

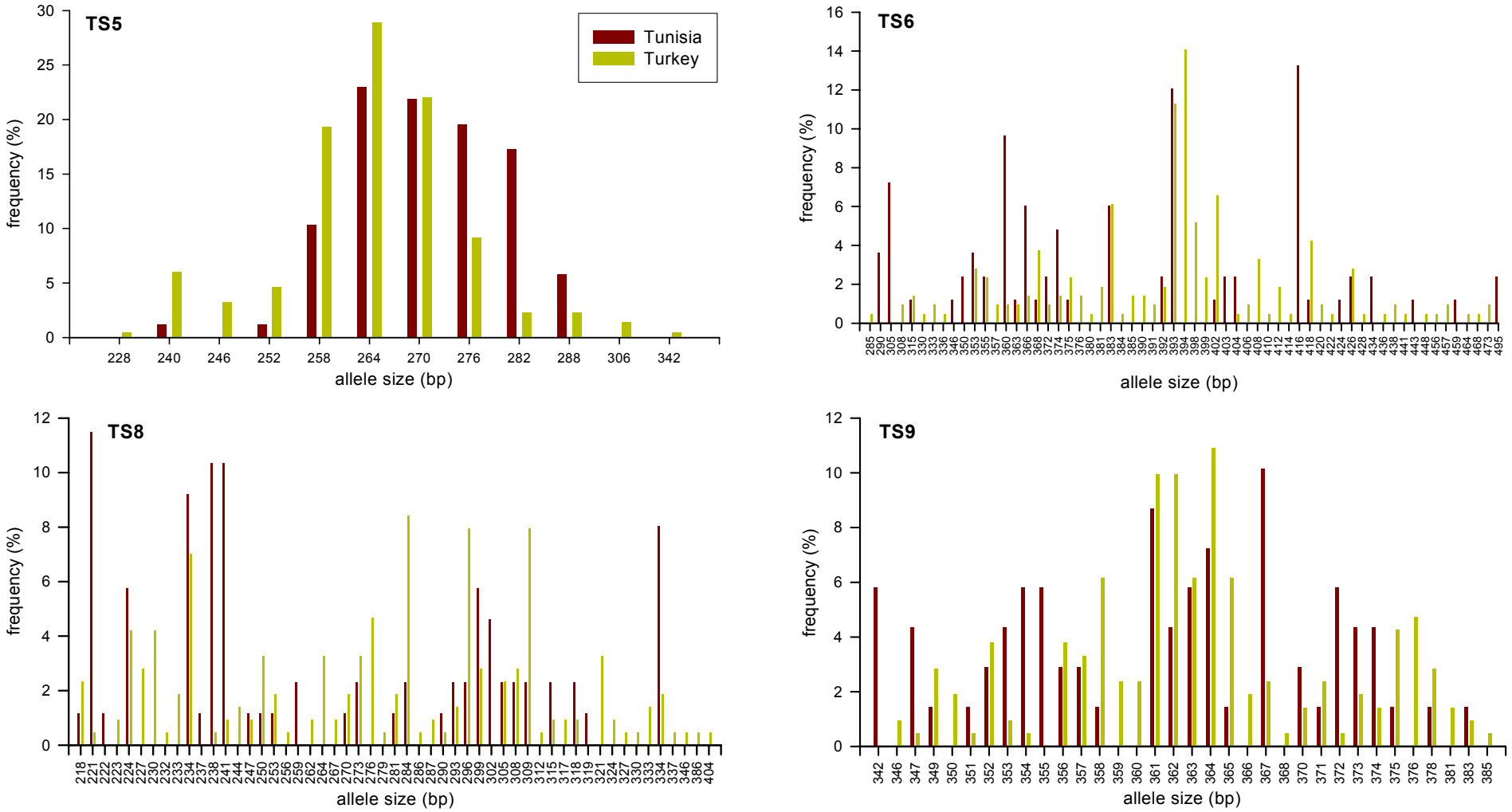
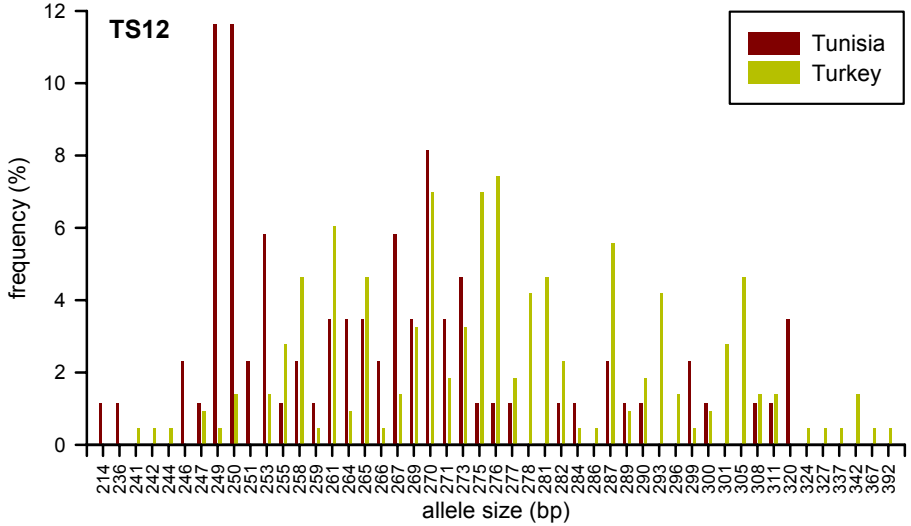
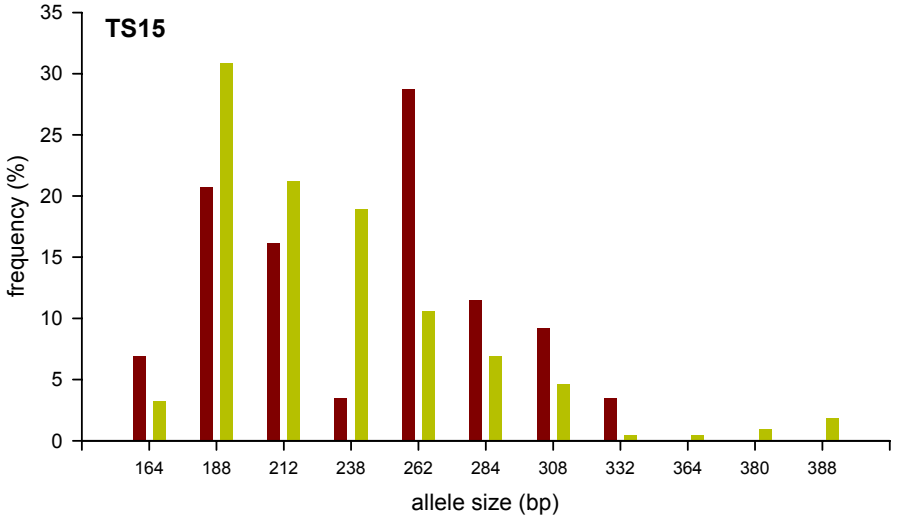
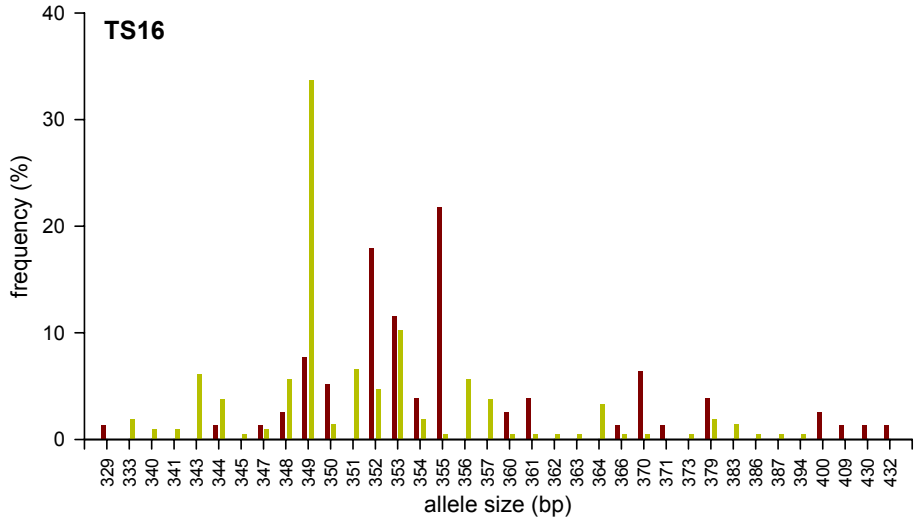


Figure 3.5. New Tunisian and Turkish allele frequencies (continued)







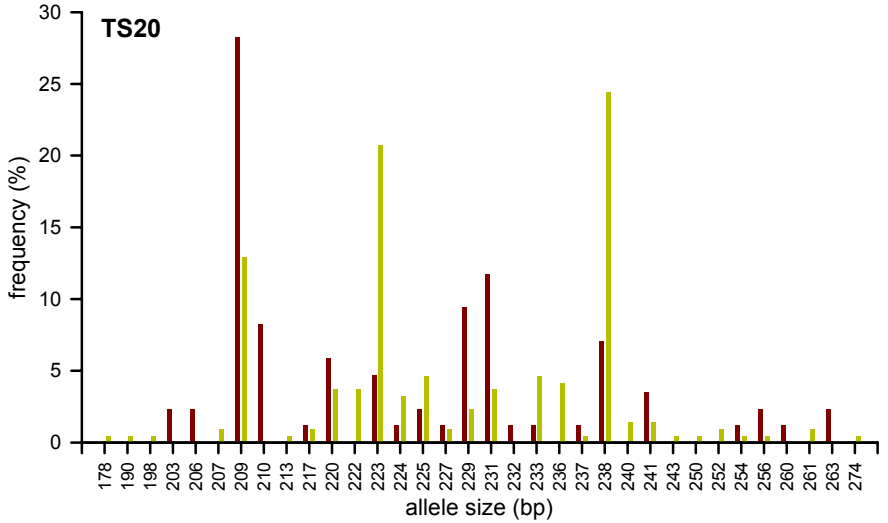
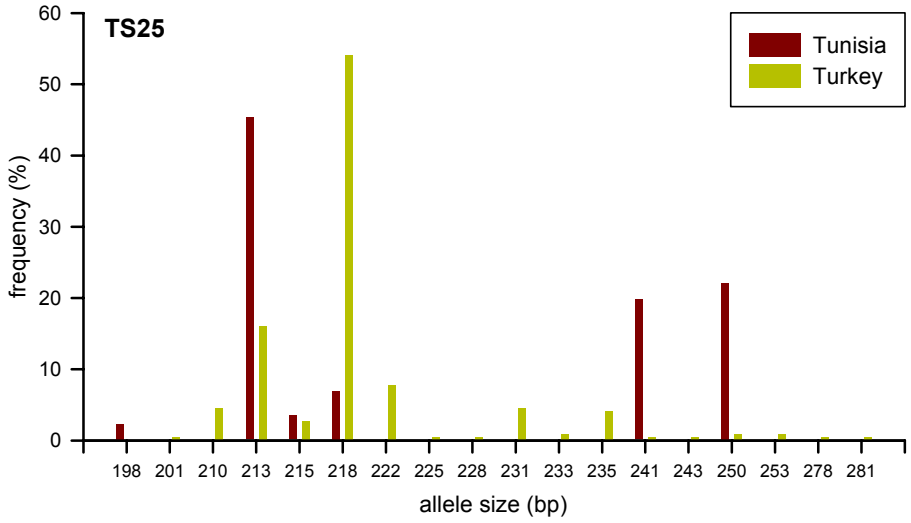


Figure 3.5. New Tunisian and Turkish allele frequencies (continued)



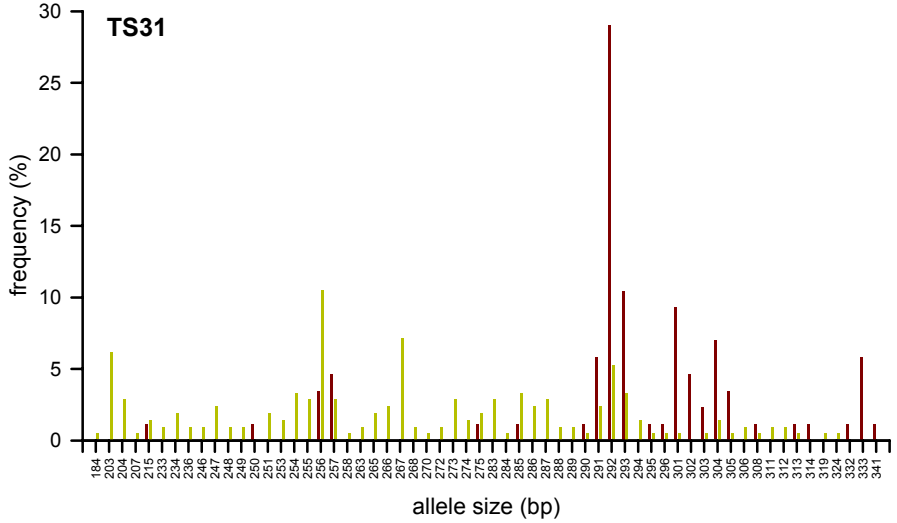


Figure 3.6. Estimated heterozygosity

(i) Initial and new analysis

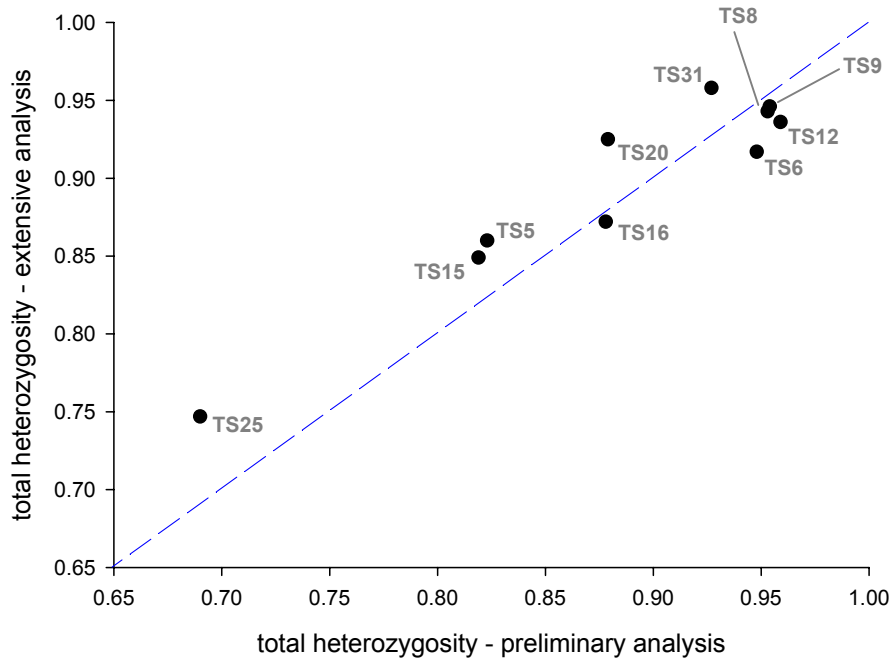
Estimated heterozygosity was compared at each of the ten loci using both the initial and new datasets. Data points plotted above the dotted blue line represented loci with increased heterozygosity in the new analysis compared to the initial study. Values were largely concordant between the two studies.

(ii) New Tunisian and Turkish populations

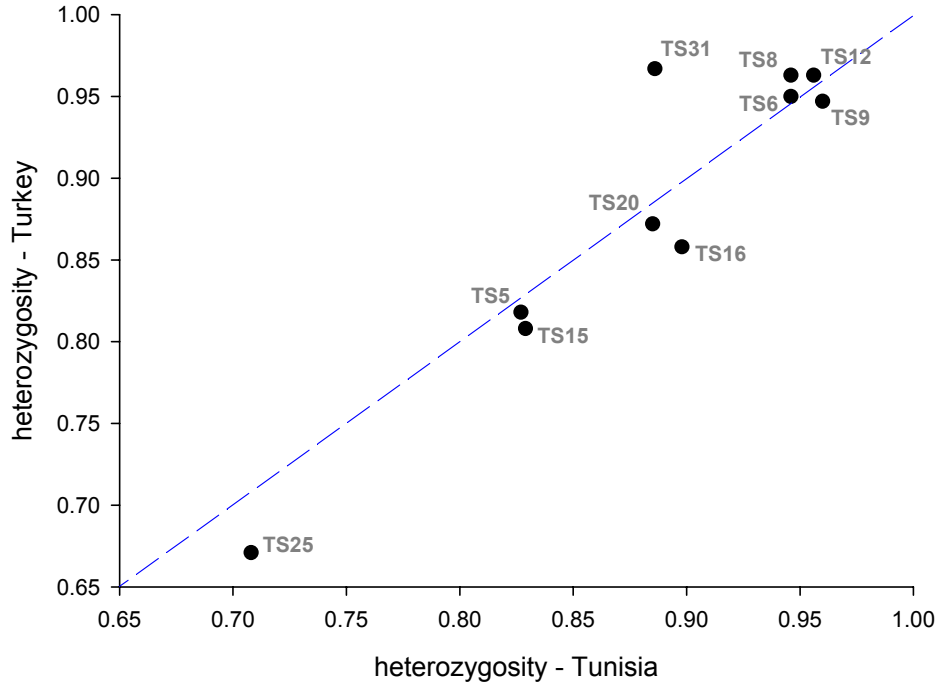
Estimated heterozygosity was compared between the new Tunisian and Turkish populations at each of the ten loci. Data points plotted above the dotted blue line represented loci with increased heterozygosity in the Tunisian population. Values were largely concordant between the two populations.

Figure 3.6. Estimated heterozygosity

(i) Initial and new analysis



(ii) New Tunisian and Turkish populations



heterozygosity for TS25 in each study (Figure 3.6.(i).) and in each population (Figure 3.6.(ii)) was related to the bias of that locus towards a single predominant allele in each population. The largest discrepancy between estimated heterozygosity values between the two populations was displayed by TS31 (Figure 3.6.(ii)).

3.3.2. Population genetic analysis

In order to confirm the presence of macro-geographical sub-structuring of populations of *T. annulata* as tentatively concluded in Chapter Two, F_{ST} values were estimated using G_{ST}' and θ across all ten markers between Tunisia and Turkey. The results of this analysis are presented in Tables 3.3. and 3.4. Values for individual markers ranged from 0.010 for TS9 to 0.221 for TS25, with a mean of 0.052 across all loci, indicating a moderate amount of differentiation. This was in general agreement with the value of 0.044 obtained during the initial study and as previously observed, a negative correlation was present between H_S and G_{ST}' (Pearson correlation co-efficient = -0.819, $p = 0.004$). The relationship of the G_{ST}' values calculated for this dataset and that presented in Chapter Two is shown graphically in Figure 3.7. where for six markers similar values were obtained using both datasets. For markers TS5, TS15 and TS31, which previously generated negative values, positive values were returned by the new dataset.

To test whether differentiation could be detected on a smaller geographical scale, populations within each country were analysed and the results are presented in Table 3.4. In addition to G_{ST}', the alternative estimator for F_{ST}, θ was calculated for all comparisons with almost identical values being returned in each case. Confirming the tentative conclusions presented in the previous chapter, a reduced amount of differentiation was observed within each country. When the Tunisian populations from Béja and El Hessiène were compared, F_{ST} was estimated at 0.012, significantly lower than the value of 0.052 measured between the countries. Interestingly, when the three farms in the village of El Hessiène were compared, G_{ST}' was slightly higher at 0.017. This indicated that the amount of differentiation observed between neighbouring sampling sites was of a similar magnitude to that observed when a distance of 100 km separated sampling sites. When the four districts in Western Turkey were compared, F_{ST} was estimated at 0.028 with a low SE for θ at 0.005. This indicated a higher amount of differentiation than observed within Tunisia, but considerably less than that detected between countries. These results support the preliminary conclusions that Tunisian and Turkish isolates do not comprise a single population of *T. annulata* and that greater parasite diversity is observed between countries than within them. This suggests a degree of genetic isolation, and therefore to test at what

Table 3.3. Indices of marker diversity and differentiation

The degree of differentiation between the new isolates representing populations in Tunisia and Turkey was determined by comparing within-sample gene diversity (H_S) with overall gene diversity (H_T'). Averaged over all ten loci, F_{ST} was estimated at 0.052, implying a moderate amount of genetic differentiation between populations, consistent with the results of the preliminary analysis (Table 2.9.).

Table 3.3. Indices of marker diversity and differentiation

Locus	H_s	H_T	H_T'	G_{ST}	G_{ST}'
TS5	0.823	0.832	0.841	0.011	0.022
TS6	0.948	0.961	0.975	0.014	0.028
TS8	0.954	0.965	0.976	0.011	0.022
TS9	0.953	0.958	0.963	0.005	0.010
TS12	0.959	0.968	0.977	0.009	0.018
TS15	0.819	0.835	0.850	0.019	0.037
TS16	0.878	0.913	0.948	0.038	0.074
TS20	0.879	0.902	0.925	0.026	0.050
TS25	0.690	0.788	0.886	0.124	0.221
TS31	0.927	0.949	0.972	0.024	0.047
Mean	0.883	0.907	0.931	0.027	0.052

H_S = within sample gene diversity, H_T = overall gene diversity,
H_T' = overall gene diversity (independent of number of samples), G_{ST} = estimator of F_{ST}
G_{ST}' = estimator of F_{ST} (independent of number of samples)

Table 3.4. Population differentiation

Standard measurements of population differentiation were made between the new isolates representing (a) the countries of Tunisia and Turkey and (b) sampling sites within each country. G_{ST}' and θ were used to estimate population differentiation with the standard error for θ calculated in order to assess variance between loci.

Table 3.4. Population differentiation

Comparison	n	F_{ST}		
		G_{ST}'	θ	θ SE
Tunisia & Turkey	305	0.052	0.052	0.019
Tunisia				
El Hessiène & Béja	71	0.012	0.012	0.007
El Hessiène (Béchir, Hassine & Salah)	44	0.017	0.018	0.012
Turkey				
Akçaova, Aydın, Incirlova & Nazilli	201	0.028	0.028	0.005

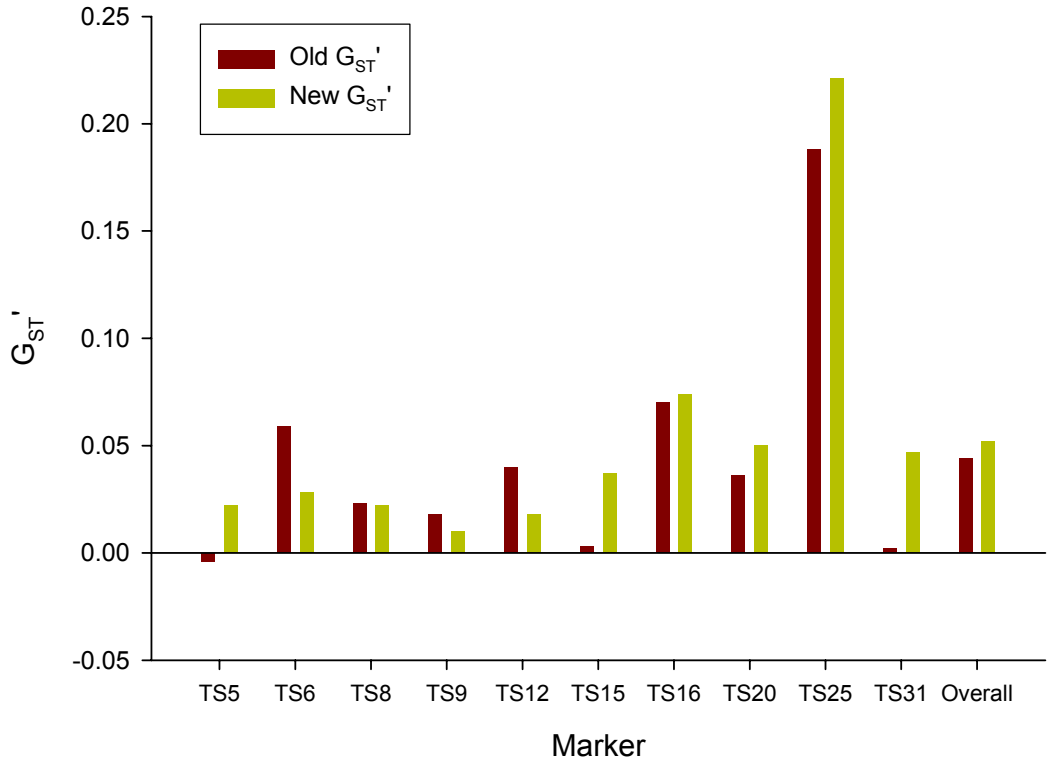
n = number of samples,

G_{ST}' and θ are estimators of F_{ST} (a measurement of differentiation), SE = standard error

Figure 3.7. Comparison of F_{ST} estimators from initial and new study

Values for the F_{ST} estimator, G_{ST}', were compared for each of the ten marker loci using the initial and new sample collections. A broadly similar pattern was observed across the majority of loci, with overall values agreeing closely between the two studies.

Figure 3.7. Comparison of F_{ST} estimators from initial and new study



resolution putative populations resolve towards linkage equilibrium, the standard index of association (I^S_A) was calculated across nested datasets representing: (1) the entire population, (2) individual sampling areas and (3) individual farms and villages. The Turkish villages of Sariköy, Osmanbuku and Sümer Mah were included in the analysis because they were each represented by a collection of greater than ten isolates; villages represented by ten or less isolates were not analysed independently. Using the dataset based solely on the predominant allele at each locus, the results of this analysis together with heterozygosity data are presented in Table 3.5. The standard index of association (I^S_A), mismatch variance (V_D) and the 95 % confidence limit for (L) are explained in Section 2.2.4. Briefly, I^S_A gives a quantitative value for the level of linkage disequilibrium - the higher the value of I^S_A, the greater the level of linkage disequilibrium, while a value close to zero indicates linkage equilibrium. In addition, linkage disequilibrium is qualitatively indicated by the relative values of V_D and L. When $V_D > L$, the null hypothesis of linkage equilibrium is disproved and linkage disequilibrium is indicated. Similar to the findings in the initial study, linkage disequilibrium (LD), indicated by $V_D > L$ was observed when samples from Tunisia and Turkey were pooled. The I^S_A was calculated as 0.0187, compared to 0.0120 from the previous study thus indicating a higher level of LD. In Tunisia, linkage equilibrium was detectable within two of the three farms at El Hessiène and within the Béja population as highlighted in Table 3.5. The 13 isolates from the Hassine farm in El Hessiène displayed a low positive I^S_A value of 0.0500. This indicated a non-random association of alleles, with particular combinations of alleles from different loci found more frequently than would be expected by chance. When the Hassine population was removed from the analysis, the rest of the Tunisian dataset reverted to linkage equilibrium, with a low I^S_A value of 0.0059 (data not shown). Consequently, LD was not detected between the Béja and El Hessiène, indicating that the Tunisian population may, in general, be described as panmictic. LD within the Hassine farm may be explained by several hypotheses, which are discussed in Section 3.4.1. Within Turkey, LD was detected when isolates from all four districts were pooled; this was associated with an increased I^S_A value of 0.0228 in comparison to the entire dataset (Turkey plus Tunisia). With the exception of Incirlova, I^S_A values above 0.0200 were obtained when each district was analysed independently, indicating the presence of LD in the other three districts. Linkage equilibrium was indicated in the district of Incirlova and the village of Osmanbuku in Aydın district. The greatest level of LD was observed in the district of Nazilli and this was, in part, due to eleven samples from the village of Sümer Mah. As presented in Table 3.5., a low level of polymorphism was indicated by a mean of only 3.00 alleles per locus. Consequently, many of these samples shared a high proportion of alleles.

Table 3.5. Heterozygosity and linkage equilibrium analysis

Heterozygosity measurement and linkage analysis were conducted on the new isolates representing (a) the countries of Tunisia and Turkey and (b) sampling sites within each country. The 'mean number of genotypes per isolate' was calculated as the mean value for the number of alleles detected at each of the ten loci. Structured combinations of populations were pooled to test for linkage disequilibrium. The standard index of association (I^S_A) provided a quantitative measurement of association of alleles. Variance of mismatch values (V_D) were compared to values of L (the upper confidence limits of Monte Carlo simulations and parametric tests), and where $V_D > L$, linkage disequilibrium (LD) was indicated. When $L > V_D$ the null hypothesis of linkage equilibrium (LE) was not disproved. Populations exhibiting linkage equilibrium are highlighted.

Table 3.5. Heterozygosity and linkage equilibrium analysis

Comparison	n	H_e	H_e SD	no. alleles	no. alleles SD	I^S_A	V_D	L_{para}	L_{MC}	Linkage
Tunisia & Turkey	305	0.9019	0.0234	36.80	18.81	0.0187	0.9582	0.8351	0.8341	LD
Tunisia	87	0.8841	0.0249	20.00	9.52	0.0102	1.0599	1.0143	1.0204	LD
El Hessiène & Béja	71	0.8826	0.0262	18.60	8.68	0.0125	1.0692	1.0111	1.0096	LD
El Hessiène (3 farms)	44	0.8717	0.0257	14.80	6.00	0.0119	1.1629	1.1414	1.1545	LD
Béchir	16	0.8879	0.0258	8.70	2.54	**-0.0168**	0.7798	1.1328	1.1047	**LE**
Hassine	13	0.8439	0.0348	6.50	1.72	0.0500	1.6818	1.4518	1.4221	LD
Salah	15	0.8478	0.0337	7.50	2.37	**0.0009**	1.2650	1.5603	1.5342	**LE**
Béja	27	0.8873	0.0263	11.80	4.52	**0.0115**	1.0054	1.0198	1.0168	**LE**
Turkey	218	0.8818	0.0305	32.60	15.95	0.0197	1.1145	0.9758	0.9769	LD
Akçaova, Aydın, Incirlova & Nazilli	201	0.8784	0.0312	31.50	15.02	0.0228	1.1659	1.0003	1.0037	LD
Akçaova	96	0.8497	0.0387	23.30	10.47	0.0252	1.3861	1.2038	1.2084	LD
Sariköy	52	0.8003	0.0459	13.60	5.40	0.0269	1.7139	1.5275	1.5192	LD
Aydın	37	0.9065	0.0230	16.40	5.72	0.0516	1.1513	0.8610	0.8656	LD
Osmanbuku	12	0.8894	0.0398	8.40	2.27	**0.0072**	0.8963	1.0828	1.1117	**LE**
Incirlova	30	0.8682	0.0328	13.00	5.87	**0.0097**	1.1111	1.1480	1.1444	**LE**
Nazilli	38	0.8421	0.0289	13.00	5.16	0.1262	2.6514	1.3743	1.3893	LD
Sümer Mah	11	0.4891	0.0690	3.00	1.25	0.0564	3.2209	2.9993	3.1468	LD

n = number of samples, H_e = estimated heterozygosity, SD = standard deviation, no. alleles = mean number of alleles identified across ten loci

I^S_A = standard index of association, V_D = mismatch variance (linkage analysis), LD = linkage disequilibrium, LE = linkage equilibrium,

L_{MC} and L_{PARA} = upper 95 % confidence limits of Monte Carlo simulation and parametric tests respectively (linkage analysis)

This was reflected in the estimated heterozygosity of 0.4891, which was considerably below the normal range of 0.8 – 0.9 that was observed in other villages and districts and suggested that the isolates from Sümer Mah were highly related to each other. When the Sümer Mah samples were removed from the analysis, the Nazilli population remained in linkage disequilibrium, although the I^S_A value dropped from 0.1262 to 0.0628 (data not shown). Therefore, in the Nazilli district, LD could be detected across the other isolates. The stronger LD (i.e. higher I^S_A) detected in Nazilli as a whole compared to the Sümer Mah and non-Sümer Mah isolates indicated a degree of sub-structuring between sampling areas within the district itself and this is also discussed in Section 3.4.1. To further investigate the distribution of genotypes between areas and to determine to what extent MLGs from the same area were related, similarity analysis was undertaken.

3.3.3. Similarity analysis

MLGs representing all the isolates from Tunisia and Turkey were compared to test if sub-structuring could be detected between countries. The allelic data contained across all ten loci was used to a construct a difference matrix and the results projected onto explanatory axes using the technique of principal component analysis (PCA). This technique was used in preference to dendrogram construction since it could better represent the continuous nature of the variation present in the dataset. The two principal axes generated by the PCA analysis are presented in Figure 3.8. The Tunisian isolates clustered almost exclusively on the rightmost quadrants of the diagram, while the Turkish isolates lay over a wider area to the left. Despite the presence of an interface region and a number of outliers, a trend was clearly evident – the primary axis discriminated between isolates from either country. In this analysis, the first and second axes explained 25 % and 18 % of the variation, and therefore the diagram represents almost half the amount of variation observed across the dataset. In addition, the contents of the difference matrix were used to construct a dendrogram, which also indicated geographical sub-structuring, although it was unstable and many outliers were present (data not shown).

To investigate whether sub-structuring could be identified between regions within each country further analyses were performed, the results of which are presented in Figure 3.9. The PCA for the Tunisian isolates was unable to clearly separate the three farms in El Hessiène (Béchir, Hassine and Salah) and the site at Béja (Figure 3.9.(i)). However, it was observed that the majority of the 27 isolates from Béja clustered in the upper two quadrants while most of the isolates from the Salah farm were found in the lower two quadrants. Considerable linkage disequilibrium was previously identified in the Hassine population,

Figure 3.8. PCA of Tunisian and Turkish isolates

Principal component analysis (PCA) was performed on the multi-locus genotype data representing the new Tunisian and Turkish populations. The two principal axes generated by this analysis are presented opposite, demonstrating a degree of sub-structuring between isolates from each country. The proportion of the variation in the dataset explained by each axis is indicated in parenthesis.

Figure 3.8. PCA of Tunisian and Turkish isolates

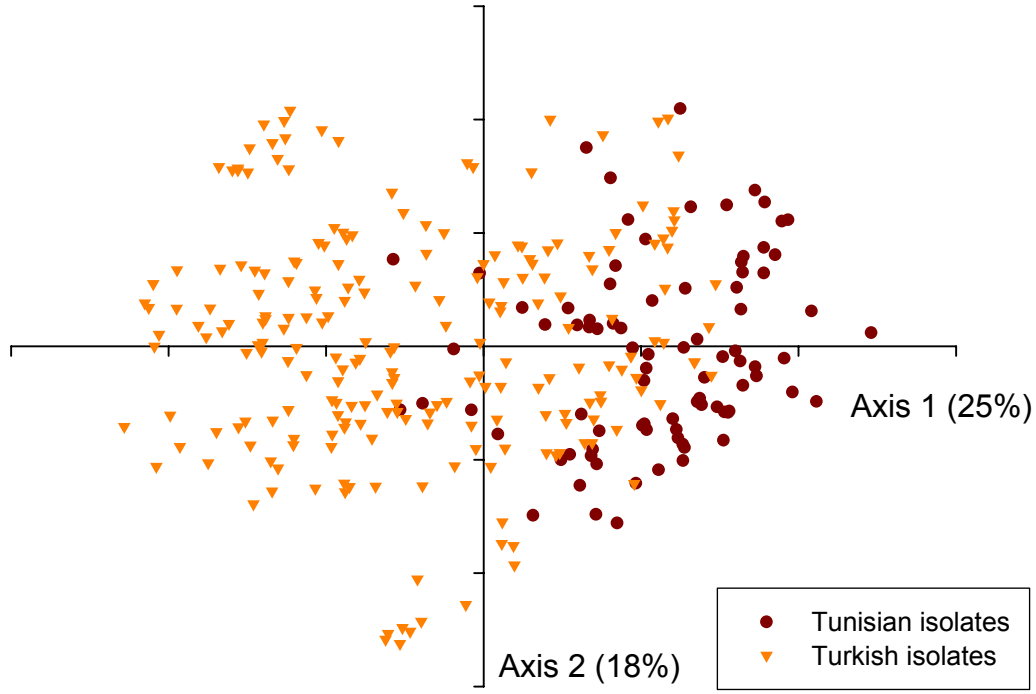


Figure 3.9. Genotypes within individual countries with respect to sampling site

PCA analysis was performed on the multi-locus genotype datasets representing each of the new sample collections from Tunisia and Turkey. The two principal axes generated by each of these analyses are presented opposite. Data points representing isolates are colour-coded to indicate their place of origin.

(i) Tunisian sites

The sampling sites in El Hessiène village and Béja are indicated. A cluster corresponding to six isolates from Hassine farm in El Hessiène is highlighted.

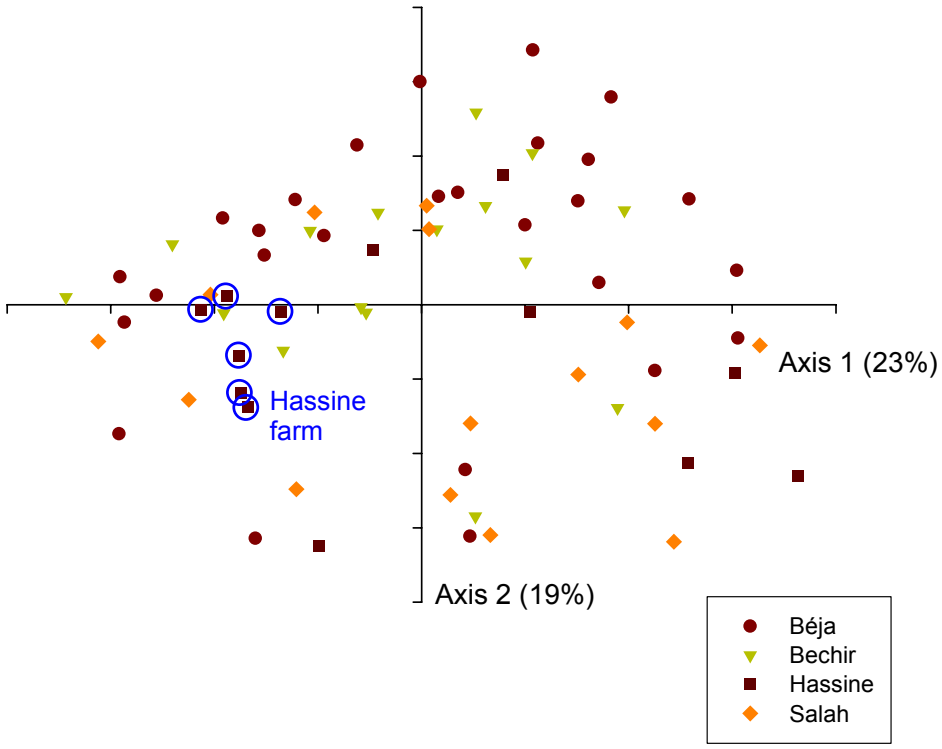
(ii) Turkish sites

The four districts in western Turkey from where parasites were isolated are indicated. A cluster corresponding to eleven isolates from the village of Sümer Mah in Nazilli is highlighted. This cluster includes a single isolate from Ocakli village, also in Nazilli.

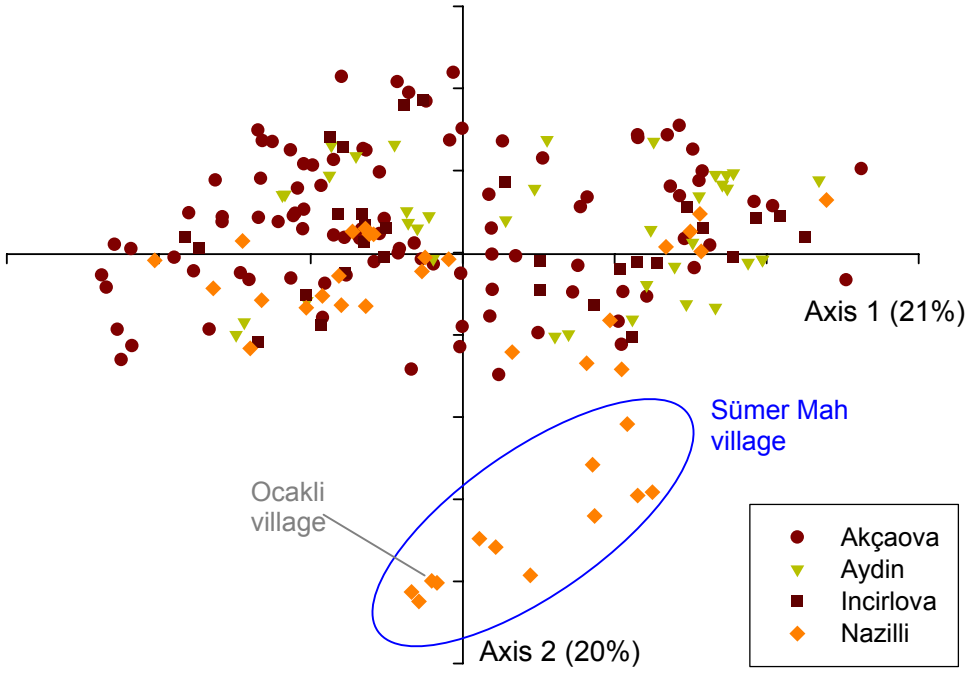
.

Figure 3.9. Genotypes within individual countries with respect to sampling site

(i) Tunisian sites



(ii) Turkish sites



indicating a high level of relatedness among isolates from this farm. This corresponded to a cluster of six isolates, which are circled in Figure 3.9.(i) and which accounted for half of the isolates from that site. Similarity analysis was performed on the isolates from each country to test whether samples from the same district or village were related to each other. In Tunisia pairs of samples sharing 40 % or more alleles were identified and are presented in Table 3.6. In the Tunisian population, no two isolates possessed an identical MLG, with a maximum of five alleles shared between any pair-wise combination of isolates. Seven isolates from Hassine (HAS001, 002, 004, 006, 010, 012 and 013) were demonstrated to share up to five alleles in their MLGs, which was consistent with the finding of linkage disequilibrium in the population from this farm. Six of these seven isolates represented the Hassine cluster, which was highlighted in Figure 3.9.(i). These and other isolates from Hassine were also shown to share a proportion of alleles with isolates from the other two farms in El Hessiène and from Béja. In comparison to Tunisia, in Turkey generally a higher level of pair-wise similarity was identified and pairs of samples sharing 60 % or more alleles are presented in Table 3.7. A marked trend was evident with closely related samples coming from the same district (Table 3.7., almost identical colouration of matching isolates). In other words, when samples shared six or more alleles, it was highly likely that they were derived from the same district. This finding is illustrated in Figure 3.10., where the pair-wise similarity matches are presented in classes, representing the percentage of shared alleles. Figure 3.10.(i) shows that a large number of pair-wise combinations (almost 8,000) shared either one or two alleles (i.e. 10 – 20 %), whereas a relatively small number of combinations shared between 60 % and 100 % of alleles. The proportion of matches coming from the same district is presented in Figure 3.10.(ii). Where only one or two alleles matched, there was around a 30 % chance the pair of samples came from the same district. In contrast when between 60 % and 100 % of alleles matched there was a 92 % chance they came from the same district. In the latter class, there were only three exceptions - three samples from Incirlova showed 60 % similarity with isolates from Akçaova (Table 3.7.). A degree of structure was also evident in the PCA analysis representing genotypes across the four districts in Western Turkey (Figure 3.9.(ii).). An independent cluster is clearly seen, which corresponds to all eleven isolates from Sümer Mah village together with a single isolate from Ocakli village (Table 3.1.), which is also in the Nazilli district. The population genetic analysis had highlighted the Sümer Mah population as having low heterozygosity and a high I^S_A value, indicating it comprised particularly closely related individuals. Although there was no clear separation between districts, a proportion of isolates from the district of Akçaova clustered loosely on the leftmost quadrants. These results suggested that there is a degree

Table 3.6. Matching Tunisian multi-locus genotypes

Pair-wise similarity analysis was performed on the multi-locus genotypes representing the new isolates from Tunisia. Pairs of samples sharing more than 40 % of alleles were identified and ranked in descending order of similarity. In some instances, where a null allele was identified at a particular locus in one of the two samples (i.e. a failure to amplify PCR product) this locus was excluded from the comparison, resulting in fewer than ten loci being compared.

Table 3.6. Matching Tunisian multi-locus genotypes

Sample 1	Sample 2	Number of loci compared	Number of identical alleles	Similarity
SAL005	SAL006	7	4	57.14%
BCR004	UNK004	9	5	55.56%
BCR013	SAL008	9	5	"
BEJ014	UNK016	9	5	"
BEJ026	SAL008	9	5	"
HAS004	HAS012	9	5	"
HAS006	UNK013	9	5	"
SAL002	SAL009	9	5	"
BCR013	BEJ002	10	5	50.00%
BEJ017	BEJ026	10	5	"
HAS001	HAS013	10	5	"
HAS002	HAS010	10	5	"
HAS006	SAL010	10	5	"
HAS007	HAS013	10	5	"
BCR016	HAS001	8	4	"
BCR016	SAL010	8	4	"
BCR016	UNK014	8	4	"
BEJ003	BEJ011	8	4	"
BEJ013	SAL001	8	4	"
BEJ027	SAL003	8	4	"
HAS003	SAL003	8	4	"
SAL003	SAL005	8	4	"
BCR001	SAL002	9	4	44.44%
BCR003	BEJ001	9	4	"
BCR003	UNK008	9	4	"
BCR003	UNK016	9	4	"
BEJ002	SAL008	9	4	"
BEJ005	SAL008	9	4	"
BEJ010	BEJ017	9	4	"
BEJ010	BEJ023	9	4	"
BEJ010	BEJ026	9	4	"
BEJ010	BEJ027	9	4	"
BEJ013	UNK012	9	4	"
BEJ015	UNK007	9	4	"
BEJ024	BEJ025	9	4	"
BEJ027	HAS003	9	4	"
BEJ027	SAL002	9	4	"
HAS001	UNK013	9	4	"
HAS002	SAL005	9	4	"
HAS003	SAL010	9	4	"
HAS006	UNK012	9	4	"
HAS008	SAL002	9	4	"
SAL008	UNK013	9	4	"
SAL010	UNK013	9	4	"
UNK001	UNK014	9	4	"
UNK005	UNK012	9	4	"
BCR016	UNK013	7	3	42.86%
HAS011	SAL014	7	3	"
SAL003	SAL006	7	3	"

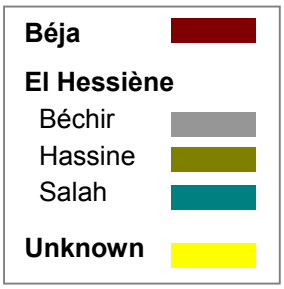


Table 3.7. Matching Turkish multi-locus genotypes

Pair-wise similarity analysis was performed on the multi-locus genotypes representing the new isolates from Turkey. Pairs of samples sharing 60 % or more alleles were identified and ranked in descending order of similarity. In some instances, where a null allele was identified at a particular locus in one of the two samples (i.e. a failure to amplify PCR product) this locus was excluded from the comparison, resulting in fewer than ten loci being compared.

Table 3.7. Matching Turkish multi-locus genotypes

Sample 1	Sample 2	Number of loci compared	Number of identical alleles	Similarity
AYD008	AYD009	10	10	100.00 %
AYD002	AYD005	10	9	90.00 %
NAZ001	NAZ005	10	9	"
NAZ005	NAZ007	10	9	"
AYD015	AYD016	10	8	80.00 %
NAZ001	NAZ007	10	8	"
NAZ007	NAZ010	10	8	"
AKC016	AKC037	9	7	77.78 %
AKC046	AKC050	10	7	70.00 %
AYD010	AYD011	10	7	"
AYD036	AYD037	10	7	"
NAZ001	NAZ003	10	7	"
NAZ001	NAZ006	10	7	"
NAZ001	NAZ012	10	7	"
NAZ002	NAZ004	10	7	"
NAZ003	NAZ005	10	7	"
NAZ005	NAZ006	10	7	"
NAZ005	NAZ010	10	7	"
NAZ005	NAZ012	10	7	"
NAZ006	NAZ012	10	7	"
NAZ009	NAZ010	10	7	"
NAZ023	NAZ024	10	7	"
NAZ024	NAZ028	10	7	"
AKC034	INC011	9	6	66.67 %
AKC040	AKC058	8	5	62.50 %
AKC054	AKC058	8	5	"
AKC007	AKC040	10	6	60.00 %
AKC017	AKC020	10	6	"
AKC017	AKC037	10	6	"
AKC020	AKC046	10	6	"
AKC021	AKC057	10	6	"
AKC029	INC030	10	6	"
AKC034	AKC043	10	6	"
AKC034	AKC044	10	6	"
AKC036	AKC044	10	6	"
AKC038	AKC049	10	6	"
AKC051	INC005	10	6	"
NAZ001	NAZ010	10	6	"
NAZ002	NAZ012	10	6	"
NAZ003	NAZ006	10	6	"
NAZ003	NAZ007	10	6	"
NAZ003	NAZ012	10	6	"
NAZ004	NAZ007	10	6	"
NAZ004	NAZ009	10	6	"
NAZ006	NAZ007	10	6	"
NAZ006	NAZ010	10	6	"
NAZ007	NAZ009	10	6	"
NAZ007	NAZ012	10	6	"
NAZ008	NAZ009	10	6	"
NAZ023	NAZ028	10	6	"

Akçaova
Aydın
Incirlova
Nazilli

Figure 3.10. Matching multi-locus genotypes in Turkey

The 218 isolates comprising the new Turkish sample collection were subjected to pair-wise similarity analysis. Combinations of isolates matched at between one and ten loci.

(i) Total number of pair-wise matches identified across four Turkish districts

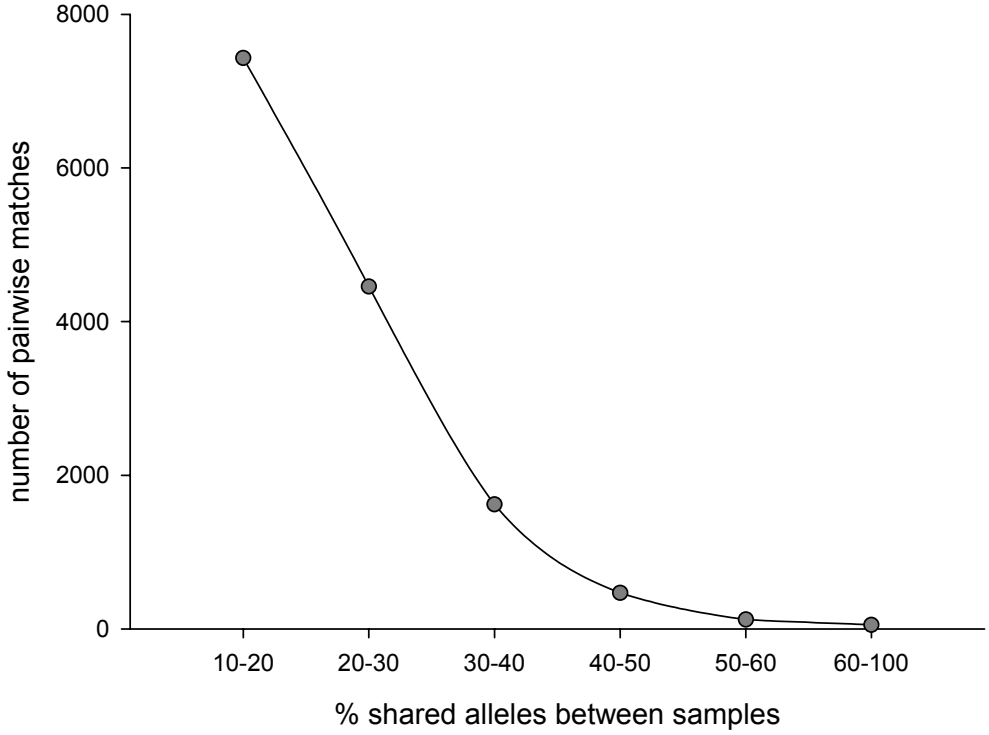
The percentage of shared alleles was used to classify similarity matches and the number of pair-wise matches in each class was calculated.

(ii) Proportion of pair-wise matches identified from the same Turkish district

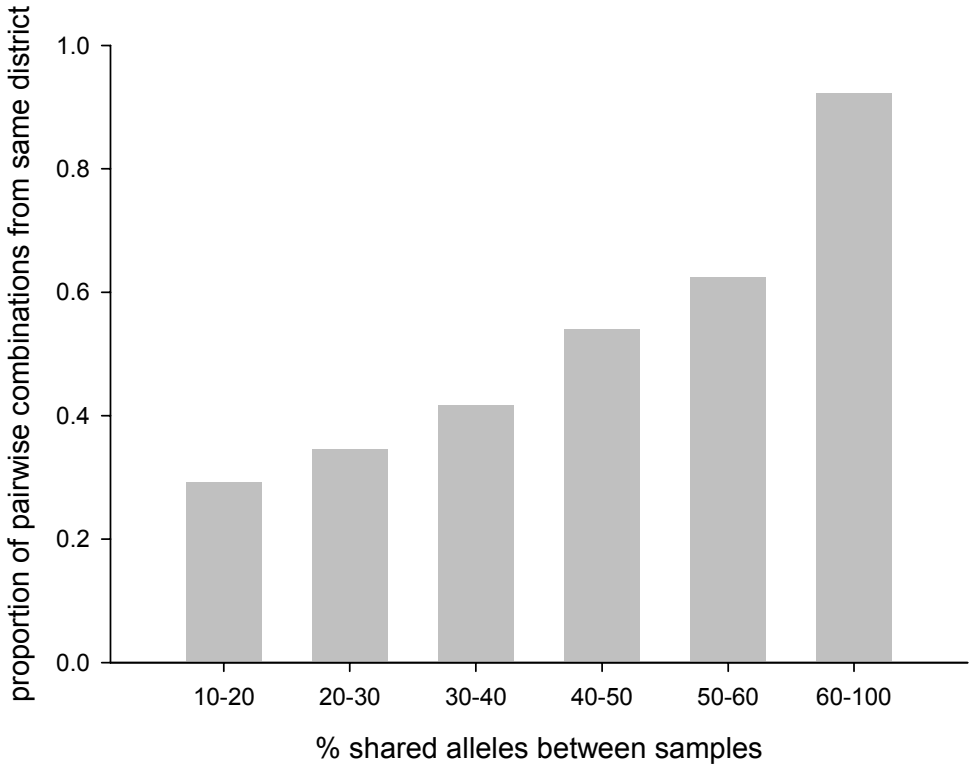
For each class of pair-wise similarity matches, the proportion of matches consisting of isolates from the same district was calculated. This demonstrated that the more similarity a pair of isolates exhibits, the more likely they are to originate from the same district.

Figure 3.10. Matching multi-locus genotypes in Turkey

(i) Total number of pair-wise matches identified across four Turkish districts



(ii) Proportion of pair-wise matches identified from the same Turkish district



of sub-structuring and so an alternate method of analysing the clustering was utilised to further investigate whether micro-geographical sub-structuring could be detected in either country. Allele similarity matrices of isolates from each country were generated and used to construct dendrograms using the method described in Section 2.2.4. Dendrograms representing Tunisian and Turkish isolates are presented in Figure 3.11. and Figure 3.12. respectively. For clarity, the earliest branches on the Tunisian dendrogram are denoted by the letters V through to Z. A limited degree of sub-structure is evident between the sampling sites in Tunisia, with isolates from Béja being over-represented in the earliest three branches (V, W and X). However, isolates from both Béja and El Hessiène are represented in each branch and there is limited association between isolates from the same farms in El Hessiène. Interestingly, the Hassine cluster, highlighted in the corresponding PCA diagram (Figure 3.9.(i)) is indistinct using the dendrogram method of illustration. These six Hassine isolates are instead grouped as pairs and are located in branches X and Y. This underscores the ability of PCA analysis to illustrate complex relationships between isolates which dendrograms may be incapable of revealing. However, dendrograms do have the advantage of readily indicating the close relationships between individual pairs of samples, since trees are generated from the 'leaves' down. The dendrogram of Turkish isolates presented in Figure 3.12. was colour-coded to represent villages that contributed nine or more isolates to the dataset; villages with less than nine isolates were denoted as 'other'. The earliest branches in the dendrogram are denoted by letters A through to I. In general, there is a tendency for isolates from the same village to cluster together suggesting a significant element of geographical sub-structuring. This is particularly evident in the isolates from Sümer Mah, which are both highly distinct from the other clusters and cluster together indicating significant similarity, thus supporting the results of the earlier PCA analysis (Figure 3.9.(ii)). Moreover, additional isolates from other regions also cluster with the Sümer Mah group in branch E of the dendrogram. It can be seen from the dendrogram that isolates from Sariköy village cluster predominantly in branch I, although other sites from within Akçaova appear scattered throughout every other branch except E. Pairs of isolates from Osmanbuku village can be seen on several branches of the dendrogram, illustrating the utility of this method to reveal pair-wise similarity. It should be stressed that although on each dendrogram, the early branching clusters appear quite distinct (Figure 3.11., V & W and Figure 3.12., A, B, C & D), possibly implying non-geographical sub-structuring, the branch lengths have been standardised, which exaggerates the distinction between these clusters and the rest of the isolates.

Figure 3.11. Dendrogram representing Tunisian sampling sites

A similarity comparison of the multi-locus genotype data representing the Tunisian isolates from Béja and El Hessiène was performed using an allele sharing co-efficient (Bowcock *et al.* 1994). The results were clustered using an unweighted arithmetic average method and used to construct a dendrogram. The dendrogram is presented as a rectangular cladogram solely to illustrate the tree topology and therefore a scale is not indicated. Isolates from Béja and the three farms in El Hessiène village were colour-coded and the earliest branches of the dendrogram were denoted by letters V through to Z. Entries marked with an asterisk (*), represent the six isolates from Hassine farm previously highlighted in Figure 3.9.(i).

Figure 3.11. Dendrogram representing Tunisian sampling sites

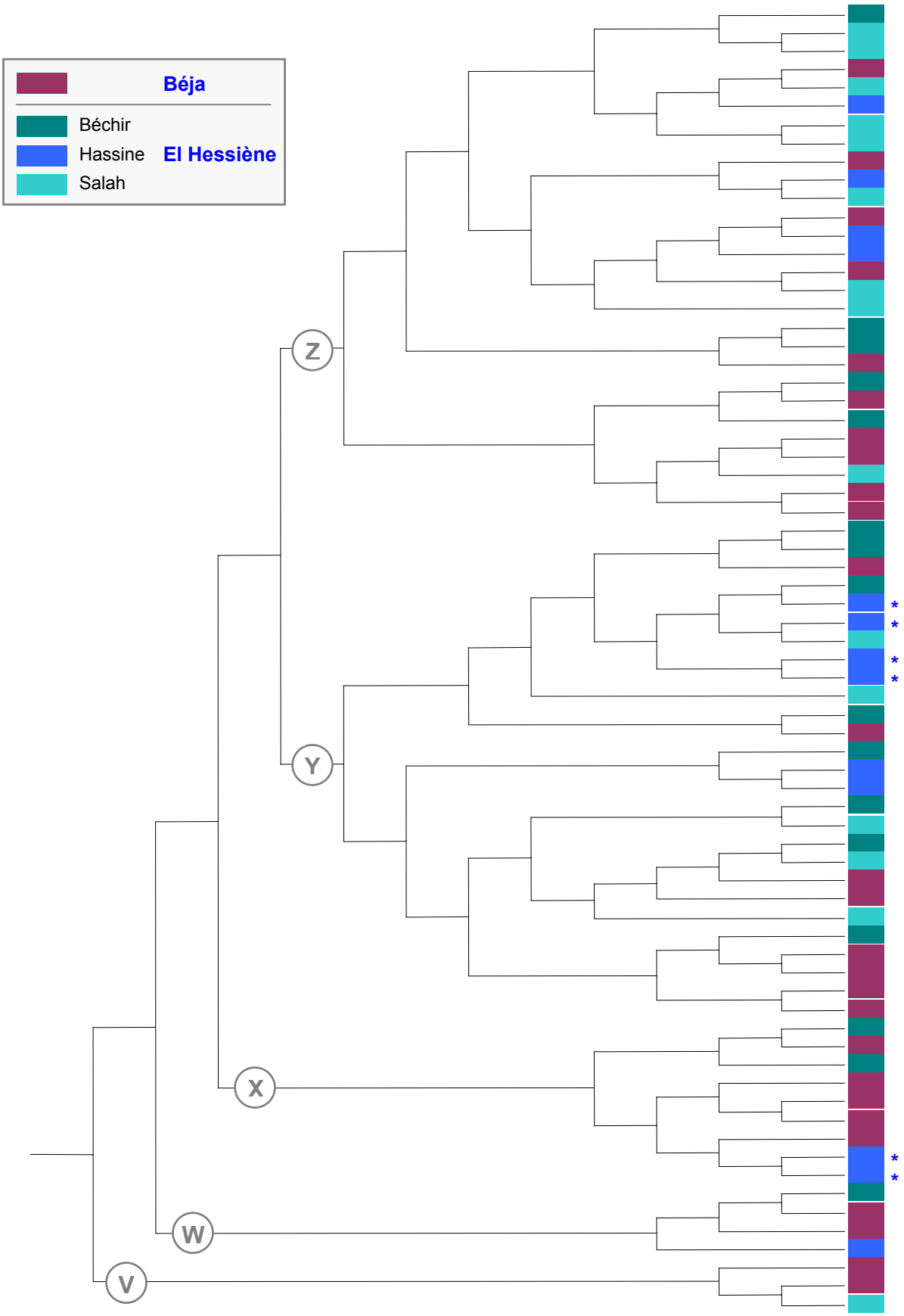
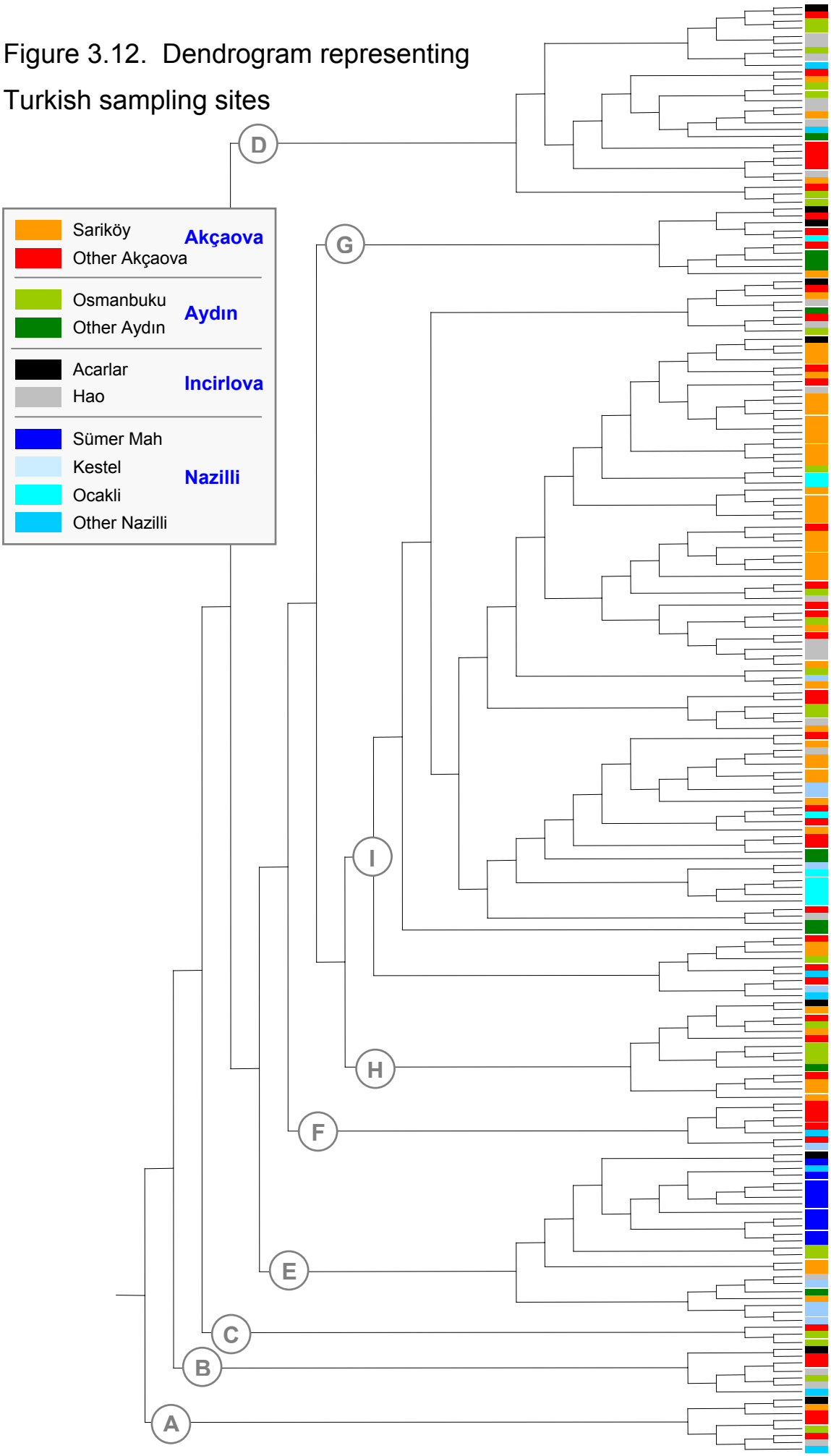


Figure 3.12. Dendrogram representing Turkish sampling sites

A similarity comparison of the multi-locus genotype data representing the isolates from the four districts in Turkey was performed using an allele sharing co-efficient (Bowcock *et al.* 1994). The results were clustered using an unweighted arithmetic average method and used to construct a dendrogram. The dendrogram was presented as a rectangular cladogram solely to illustrate the tree topology and therefore a scale was not indicated. Isolates from the largest sampling sites in each district were colour-coded and the earliest branches of the dendrogram were denoted by letters A through to I.

Figure 3.12. Dendrogram representing Turkish sampling sites



Generally, the similarity analyses indicate that there is significant sub-structuring of parasite populations between countries and an element of sub-structure within each country, which is most clearly evident in Turkey.

The blood samples used to provide parasite material were collected over the space of two years in Tunisia and in 1996, 2001 and 2003 in Turkey and thus cover more than one disease season. To investigate whether the time of sampling could explain the distribution of genotypes within each country, the PCA analysis was relabelled to denote the year of origin (Figure 3.13., these figures are based on the same MLG data as presented in Figure 3.9.). In the case of the Tunisian samples no specific trend is evident. Interestingly, the data points representing the Hassine cluster, identified in Figure 3.9.(i) are no longer distinct. This indicates that the isolates collected from Hassine farm clustered together despite being sampled across different disease seasons. In the case of the Turkish samples, a broadly similar pattern is observed in the PCA denoting year of sampling (Figure 3.13.(ii)) to the PCA denoting site of origin (Figure 3.9.(ii)). This is consistent with the fact that the samples from different areas were obtained at different points in time. It may be concluded that the analysis of differential sampling times revealed no additional trends in the distribution of genotypes.

3.3.4. Re-analysis of linkage between loci

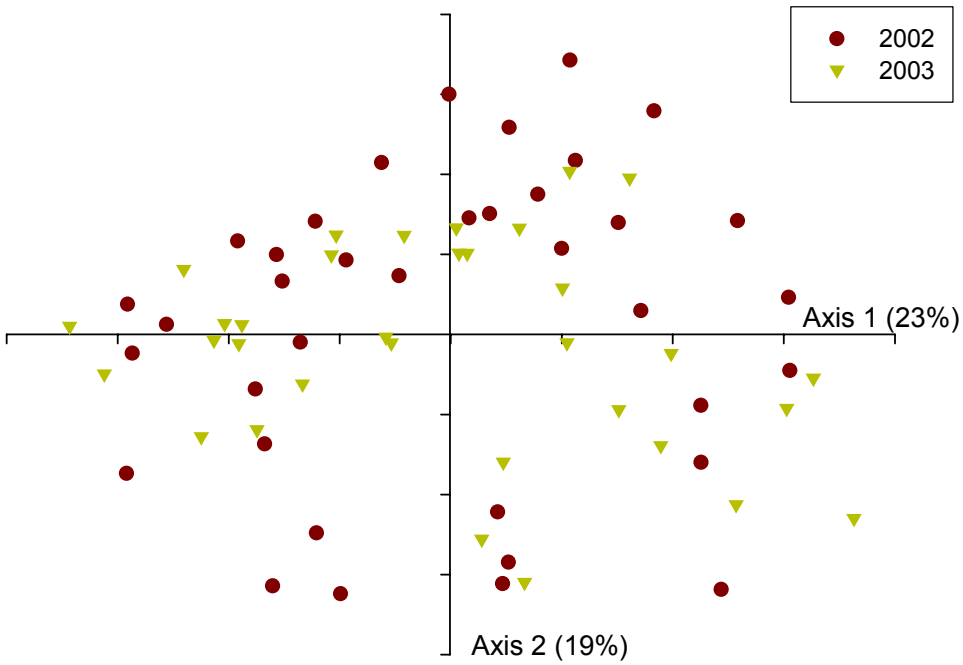
Linkage disequilibrium (LD) was identified in both Tunisian and Turkish populations in the analysis presented in Section 3.3.2. In that analysis, geographically defined sub-populations were analysed to investigate whether the site of isolation influenced genetic linkage. Following the similarity analysis presented in Section 3.3.3., a slightly different methodology was used to formally test whether a limited subset of distinct samples was responsible for LD in each country. This was intended to reveal whether an epidemic population structure existed in Tunisia and Turkey. To achieve this, a stratified approach was taken to measuring linkage within each country by successively removing the early branches from each dendrogram and re-calculating the standard index of association (I^S_A). The null hypothesis of linkage equilibrium was also re-tested and the results of each of these analyses are presented in Table 3.8. In the Tunisian collection of 71 isolates, when branches V through to X were successively removed from the analysis, there was little effect on I^S_A and LD continued to be indicated. Finally, only when the group corresponding to branch Y of the dendrogram was removed, leaving just 28 samples, did the remaining isolates exhibit LE, with an I^S_A value of -0.0034. This did not suggest that a limited subset of highly similar samples was responsible for LD within the country as a

Figure 3.13. Genotypes within individual countries with respect to year of sampling

PCA analysis was performed on the multi-locus genotype datasets representing each of the new sample collections from Tunisia **(i)** and Turkey **(ii)**. The two principal axes generated by each of these analyses are presented opposite. Data points representing isolates are colour-coded to indicate the year of sampling. These two PCA diagrams represent essentially the same information that was presented in Figure 3.9., but using a different labelling criterion.

Figure 3.13. Genotypes within individual countries with respect to year of sampling

(i) Tunisian isolates



(ii) Turkish isolates

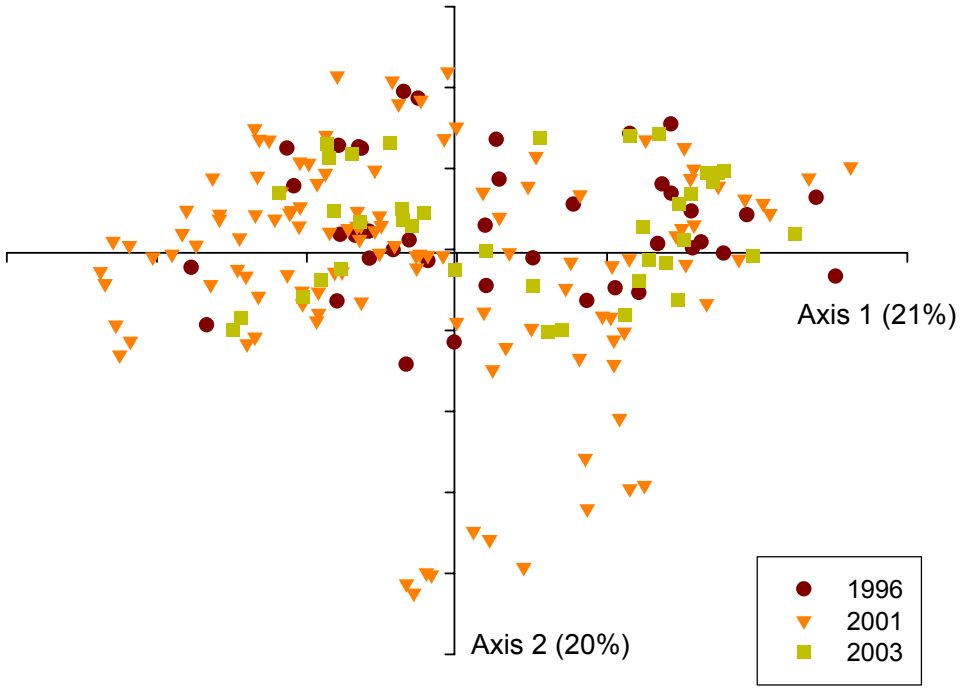


Table 3.8. Stratified linkage analysis in Tunisia and Turkey

Stratified linkage analysis was conducted on the new collections of isolates in order to test whether the early-branching clusters in the dendrograms generated for each country were responsible for masking linkage equilibrium in the rest of the population. Clusters of isolates were successively removed from each population and the remaining isolates re-tested. Groups V to Y represented the early branches in the dendrogram of Tunisian isolates (Figure 3.11.), while groups A to H represented the early branches in the dendrogram of Turkish isolates (Figure 3.12.). As before, the standard index of association (I^S_A) provided a quantitative measurement of association of alleles. Variance of mismatch values (V_D) were compared to values of L (the upper confidence limits of Monte Carlo simulations and parametric tests), and where $V_D > L$, linkage disequilibrium (LD) was indicated. When $L > V_D$ the null hypothesis of linkage equilibrium (LE) was not disproved.

Table 3.8. Stratified linkage analysis in Tunisia and Turkey

Comparison	n	I^S_A	V_D	L_{para}	L_{para} *p* value	L_{MC}	L_{MC} *p* value	Linkage
Tunisia								
El Hessiène & Béja	71	0.0125	1.0692	1.0111	*0.000*	1.0096	*0.010*	LD
Minus group V (n = 3)	68	0.0122	1.0936	1.0394	*0.000*	1.0453	*0.010*	LD
Minus group W (n = 4)	64	0.0124	1.1292	1.0779	*0.000*	1.0776	*0.010*	LD
Minus group X (n = 9)	55	0.0126	1.1867	1.1422	*0.000*	1.1571	*0.020*	LD
Minus group Y (n = 27)	28	-0.0034	1.2050	1.4242	*1.000*	1.4065	*0.630*	**LE**
Turkey								
Akçaova, Aydın, Incirlova & Nazilli	201	0.0228	1.1659	1.0003	*0.000*	1.0037	*0.010*	LD
Minus group A (n = 8)	193	0.0220	1.1877	1.0281	*0.000*	1.0269	*0.010*	LD
Minus group B (n = 7)	186	0.0212	1.2073	1.0536	*0.000*	1.0521	*0.010*	LD
Minus group C (n = 3)	183	0.0212	1.2158	1.0612	*0.000*	1.0772	*0.010*	LD
Minus group D (n = 28)	155	0.0217	1.2738	1.1183	*0.000*	1.1188	*0.010*	LD
Minus group E (n = 24)	131	0.0156	1.1900	1.1012	*0.000*	1.1019	*0.010*	LD
Minus group F (n = 7)	124	0.0136	1.1721	1.1050	*0.000*	1.1133	*0.010*	LD
Minus group G (n = 10)	114	0.0095	1.1925	1.1675	*0.002*	1.1736	*0.010*	LD
Minus group H (n = 14)	100	0.0058	1.1297	1.1469	*0.075*	1.1516	*0.120*	**LE**

n = number of samples, I^S_A = standard index of association, V_D = mismatch variance (linkage analysis), LD = linkage disequilibrium, LE = linkage equilibrium,

L_{MC} and L_{PARA} = upper 95 % confidence limits of Monte Carlo simulation and parametric tests respectively (linkage analysis)

Groups V – Y represent early branches on Tunisian isolate dendrogram (Figure 3.11.), Groups A – H represent early branches on Turkish isolate dendrogram (Figure 3.12.)

whole and thus the analysis did not provide evidence of an epidemic population structure. It was therefore concluded that geographical sub-structuring provided a better explanation of LD in Tunisia. In Turkey an I^S_A value of 0.0228 was obtained for the 201 isolates from the four districts and LD was indicated. It was not until eight of the earliest branches were removed from this dendrogram (A-H) that the remaining 100 isolates exhibited linkage equilibrium. Similar to the Tunisian isolates, an epidemic population structure was not indicated, supporting the earlier evidence of geographical sub-structuring.

An alternative explanation for LD was the physical proximity (i.e. physical linkage) of two or more markers. Although the loci of TS8 and TS9 are separated by only 38 kb on chromosome III (Figure 2.4.), no linkage was detected in the initial population studies presented in Chapter Two. In order to investigate whether physical linkage of these two loci contributed to LD in the new dataset, the TS9 locus was removed and linkage analysis was performed again using only nine loci. A comparison of the results before and after the removal of TS9 is presented in Table 3.9. I^S_A values were broadly similar between each dataset and LD or LE was consistently indicated. The only exceptions were for the El Hessiène village in Tunisia and in Sümer Mah village in the Nazilli district of Western Turkey. In contrast to the earlier analysis, LE was indicated using the nine loci and this was associated with a slight drop in I^S_A value. However, when the underlying values of V_D and L were analysed in the former ten-locus analysis, the LD was almost undetectable. Interestingly across the combined populations of both countries, a slightly increased value of I^S_A was obtained using only nine loci, 0.0200 compared to 0.0187 for ten loci. Across the comparisons, I^S_A increased slightly in the Tunisian stratified analysis and decreases slightly in the Turkish stratified analysis. Slight increases and decreases were also observed in the geographical comparisons. The slight inconsistencies between the nine-locus and ten-locus analyses were attributed to stochastic errors. It was therefore concluded that physical linkage of TS8 and TS9 did not bias the analysis towards LD and therefore the results over all ten loci were considered valid.

3.3.5. Multiplicity of infection and host phenotype

Every isolate genotyped in this study represented a mixed infection, with several alleles identified at one or more loci. The mean number of alleles for the ten loci was calculated for each isolate to provide an index value that represented the multiplicity of infection within each isolate. A summary of multiplicity of infection with respect to the area of isolation is presented in Table 3.10. Tunisian isolates possessed on average 2.51 alleles per locus, whereas Turkish isolates possessed 3.15. High standard deviation values for

Table 3.9. Linkage re-analysis omitting TS9 locus

The linkage analyses presented in Tables 3.5. and 3.8. incorporated all ten of the micro- and mini-satellite loci. To investigate whether the physical proximity of loci TS8 and TS9 (Figure 2.4.) contributed to linkage disequilibrium, the analyses were repeated omitting the TS9 locus. Results from each analysis were broadly similar, with the minor differences attributed to stochastic error, which in part was related to using a smaller dataset.

Table 3.9. Linkage re-analysis omitting TS9 locus

Comparison	n	10 loci		9 loci (TS9 omitted)	
		I^S_A	Linkage	I^S_A	Linkage
Tunisia & Turkey	305	0.0187	LD	0.0200	LD
Tunisia	87	0.0102	LD	0.0102	LD
El Hessiène & Béja	71	0.0125	LD	0.0116	LD
El Hessiène (3 farms)	44	0.0119	LD	**0.0105**	**LE**
Béchir	16	**-0.0168**	**LE**	**-0.0236**	**LE**
Hassine	13	0.0500	LD	0.0435	LD
Salah	15	**0.0009**	**LE**	**0.0049**	**LE**
Béja	27	**0.0115**	**LE**	**0.0136**	**LE**
El Hessiène & Béja	71	0.0125	LD	0.0116	LD
Minus group V (n = 3)	68	0.0122	LD	0.0109	LD
Minus group W (n = 4)	64	0.0124	LD	0.0108	LD
Minus group X (n = 9)	55	0.0126	LD	0.0111	LD
Minus group Y (n = 27)	28	**-0.0034**	**LE**	**-0.0034**	**LE**
Turkey	218	0.0197	LD	0.0210	LD
Akçaova, Aydın, Incirlova & Nazilli	201	0.0228	LD	0.0242	LD
Akçaova	96	0.0252	LD	0.0267	LD
Sariköy	52	0.0269	LD	0.0275	LD
Aydın	37	0.0516	LD	0.0509	LD
Osmanbuku	12	**0.0072**	**LE**	**-0.0101**	**LE**
Incirlova	30	**0.0097**	**LE**	**0.0091**	**LE**
Nazilli	38	0.1262	LD	0.1258	LD
Sümer Mah	11	0.0564	LD	**0.0353**	**LE**
Akçaova, Aydın, Incirlova & Nazilli	201	0.0228	LD	0.0242	LD
Minus group A (n = 8)	193	0.0220	LD	0.0231	LD
Minus group B (n = 7)	186	0.0212	LD	0.0223	LD
Minus group C (n = 3)	183	0.0212	LD	0.0224	LD
Minus group D (n = 28)	155	0.0217	LD	0.0220	LD
Minus group E (n = 24)	131	0.0156	LD	0.0163	LD
Minus group F (n = 7)	124	0.0136	LD	0.0144	LD
Minus group G (n = 10)	114	0.0095	LD	0.0100	LD
Minus group H (n = 14)	100	**0.0058**	**LE**	**0.0065**	**LE**

n = number of samples, I^S_A = standard index of association,

LD = linkage disequilibrium, LE = linkage equilibrium,

Table 3.10. Summary of multiplicity of infection by area of isolation

The multiplicity of infection was estimated for each isolate in the new sample collection by calculating the average number of alleles present at each of the ten loci; this provided an index of the number of genotypes present in each sample. A mean value for this measurement was calculated for the Tunisia and Turkey populations and for several sub-populations within each country, which were represented by nine or more isolates. Additionally, the standard deviation (SD) and the minimum and maximum values for this index of multiplicity were calculated for each group. The two sampling sites in Tunisia and the four districts in western Turkey are highlighted.

Table 3.10. Summary of multiplicity of infection by area of isolation

Sample	n	Number of alleles per locus per isolate			
		Mean	SD	Minimum	Maximum
Tunisia	87	2.51	0.76	1.25	4.60
Béja	27	2.25	0.64	1.44	3.80
El Hessiène	44	2.57	0.78	1.25	4.60
Béchir	16	2.26	0.70	1.25	4.10
Hassine	13	2.92	0.85	1.75	4.60
Salah	15	2.58	0.71	1.56	4.20
Turkey	218	3.15	1.31	1.10	6.11
Akçaova	96	3.68	1.18	1.22	6.11
Sariköy	52	4.10	1.02	1.60	6.11
Aydın	37	2.36	0.75	1.10	4.50
Osmanbuku	12	2.52	0.96	1.20	4.50
Incirlova	30	2.69	0.85	1.29	4.40
Acarlar	9	2.16	0.57	1.50	3.33
Hao	21	2.92	0.85	1.29	4.40
Nazilli	38	2.93	0.96	1.30	5.00
Sümer Mah	11	1.98	0.29	1.56	2.50
Kestel	10	3.26	0.53	2.50	4.00
Ocakli	9	4.08	0.55	3.30	5.00

n = number of samples, SD = standard deviation

each country indicated that there was a significant amount of variance in both isolate collections, with more variation in the Turkish population where maximum and minimum values of 1.10 and 6.11 were observed. In Turkey the mean value varied between 1.98 alleles per locus in Sümer Mah village (Nazilli district) to 4.10 alleles per locus in Sariköy village (Akçaova district). Estimates of multiplicity of infection were broadly similar between El Hessiène and Béja in Tunisia. The variation in multiplicity of infection between different areas may reflect differences in parasite epidemiology such as transmission intensity. This may in turn be related to variation in the level of tick infestation and / or variation in the incidence of *T. annulata* infection in ticks. However, a number of host factors may also be involved in explaining variation in multiplicity among isolates. For example, it might be predicted that multiplicity would increase with age due to increased challenge with time or that vaccination, by inducing immunity, would reduce the number of genotypes in an animal. On this basis the host variables of age, sex, breed and vaccination status were investigated. Of the 305 isolates, which constituted the Turkish collection of samples, 199 had complete data relating to each of these parameters for the host from which they were obtained. A summary of this data is presented in Table 3.11. The dataset comprised isolates in Western Turkey collected from cattle between 3 and 180 months of age, around 60 % of which were female. The majority of cattle were imported dairy breeds and around 20 % had a history of cell line vaccination. Analysis of co-variance was performed to determine if any of these variables could explain the mean number of alleles present in each isolate. First, a correlation matrix between all combinations of variables was constructed and this is presented in Table 3.12.(i). The correlation co-efficients between each pair-wise comparison are indicated, with the correlation to multiplicity of infection indicated on the bottom line of this matrix. A positive correlation was demonstrated between multiplicity of infection and the following phenotype parameters – (i) increasing age, (ii) positive vaccination status, (iii) male sex and (iv) indigenous breed. The statistical significance of each correlation was not determined and therefore the conclusions from this analysis must be interpreted with some caution. Using analysis of co-variance (Section 3.2.3.), a linear model incorporating each of these variables was constructed to predict multiplicity of infection and is presented in Table 3.12.(ii). The model treated each explanatory variable independently while assuming that a linear relationship existed between the explanatory (i.e. host variables) and the dependent variable (i.e. multiplicity of infection). Although a true linear relationship may not exist, the model may still have value in explaining a proportion of the variation in the dataset. In order to demonstrate how each variable contributed to explaining the complexity of mixed infection, the co-efficients used in this model were standardised using

Table 3.11. Distribution of host variables and multiplicity of infection in Turkish cattle

The new Turkish parasite collection included 199 samples, which were isolated from cattle whose age, breed, sex and vaccination status was known. A summary of this dataset is presented opposite, together with corresponding statistics describing multiplicity of infection. Multiplicity of infection and host age were considered continuous, quantitative variables while vaccination status, sex and breed were qualitative variables represented by two alternate states.

Table 3.11. Distribution of host variables and multiplicity of infection in Turkish cattle

(i) Summary of quantitative variables

Variable	n	Minimum	Maximum	Mean	SD
Multiplicity of infection (mean no. of alleles per locus)	199	1.10	6.11	3.19	1.15
Host age (months)	199	3	180	38.88	31.19

n = number of samples, SD = standard deviation

(ii) Summary of qualitative variables

Variable	Categories	n	%
Vaccinated	no	159	79.9
	yes	40	20.1
Sex	female	122	61.3
	male	77	38.7
Breed	imported	185	93.0
	indigenous	14	7.0

n = number of samples

Table 3.12. Co-variance of multiplicity of infection and host variables in Turkish cattle

Analysis of co-variance (ANCOVA) was undertaken on 199 Turkish isolates detailed in Table 3.11. The multiplicity of infection, estimated by the mean number of alleles per locus was correlated with the host variables of age, vaccination status, sex and breed and the co-variance is presented in blue **(i)**. This revealed the relative influence of each host variable in explaining multiplicity of infection and allowed a model equation to be devised **(ii)**. The goodness of fit (r^2) value was calculated at 0.11, indicating that only 11 % of the variability in multiplicity of infection was explained by the host variables. However, Fisher's F test was used to analyse whether the four explanatory host variables brought significant information to the model. The Fishers F value of 5.92 with four degree of freedom and 199 samples corresponded to a *p* value of 0.0002 indicating that the host variables did contribute significant information.

Table 3.12. Co-variance of multiplicity of infection and host variables in Turkish cattle

(i) Correlation matrix

Variables	Age	Unvaccinated	Vaccinated	Female	Male	Imported	Indigenous
Unvaccinated	0.210						
Vaccinated	-0.210	-1.000					
Female	0.522	0.297	-0.297				
Male	-0.522	-0.297	0.297	-1.000			
Imported	0.093	0.058	-0.058	0.225	-0.225		
Indigenous	-0.093	-0.058	0.058	-0.225	0.225	-1.000	
Multiplicity of infection (mean no. of alleles per locus)	0.070	-0.276	0.276	-0.108	0.108	-0.092	0.092

(ii) Model equation

Mean = 2.630 + 0.007 * [age] + 0.800 * [vaccinated] + 0.259 * [male] + 0.310 * [indigenous]

Goodness of fit: $r^2 = 0.11$

Fishers F test: $F_{4,199} = 5.92$, $p = 0.0002$

the ANCOVA analysis feature of the software package XLSTAT2006. These resultant standardised co-efficients represented the relative influence of each parameter in explaining the variation in the dataset. Standardised co-efficients are presented in Figure 3.14. along with their respective 95 % confidence intervals. It can be seen from this figure that increasing age and positive vaccination status were suggested as the most important factors, with the lower limit of both confidence intervals above zero. Sex and breed type explained less of the variation in multiplicity of infection between isolates, with standard correlation co-efficients not being statistically different from zero.

To further investigate the relationship between multiplicity of infection with the age of the host, isolates from the four districts in Turkey and the site at El Hessiène in Tunisia were examined independently. Linear regression analysis was performed, plotting the mean number of alleles per locus against age of the host in months and the results are presented in Figure 3.15. The cattle from the four districts in Turkey consisted mostly of individuals up to 100 months, while the population in El Hessiène was distributed principally between one and seven months of age and therefore covered a much narrower age range. In all areas, with the exception of Aydın district, a positive relationship was demonstrated. Additionally, when the dataset was treated as a whole, linear regression analysis demonstrated a statistically significant positive correlation ($p < 0.05$, data not shown). The district of Akçaova contributed the highest number of isolates, and while the gradient of the regression line was similar to that in Nazilli and Incirlova, the intersection with the y-axis is markedly higher than the other three Turkish districts. This is reflected in the mean number of alleles per locus presented in Table 3.10. These results strongly support the hypothesis that the number of *T. annulata* genotypes within an individual increases with age, both within the first disease season (i.e. El Hessiène) and over the lifetime of the host. The same regional groupings were used to investigate the influence of host sex on multiplicity of infection and the number of alleles per locus was estimated in both sexes from each of the five areas (Figure 3.16.(i)). In El Hessiène, Aydın and Akçaova the mean number of alleles per locus observed in males was very similar to that found in females. In contrast to the general trend suggested by the co-variance analysis, females showed a higher multiplicity of infection in the districts of Incirlova and Nazilli. In order to investigate whether the differences identified in these two districts were due to varying age distribution between the sexes, the dataset was re-analysed controlling for host age (Figure 3.16.(ii)). Unfortunately, in Incirolva the age of only one male was recorded and therefore no firm conclusions can be drawn between between the sexes. In Nazilli the majority of the males sampled were less than one year, causing a bias within the overall

Figure 3.14. Multiplicity of infection – standardised co-efficients of co-variance

A model to explain multiplicity of infection was constructed using the host variables of age, sex, breed and vaccination status (Table 3.11.). These co-efficients comprising this model were standardised in order to represent the relative influence of each of the variables in explaining the variation between isolates. The standardised values together with their respective 95 % confidence intervals are presented opposite.

Figure 3.14. Multiplicity of infection – standardised co-efficients of co-variance

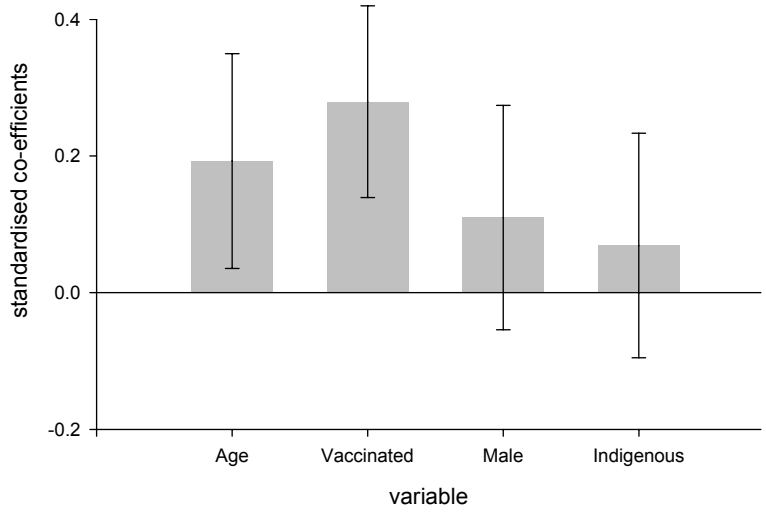
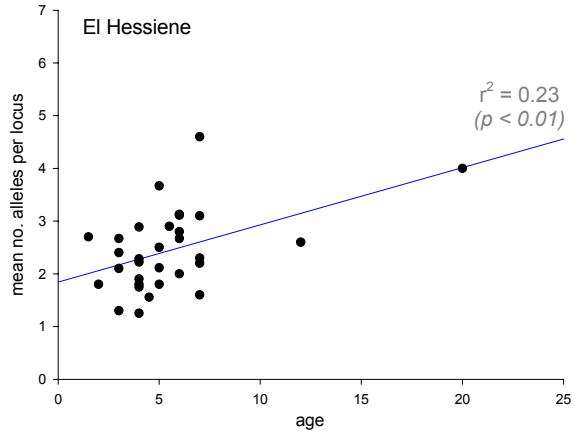


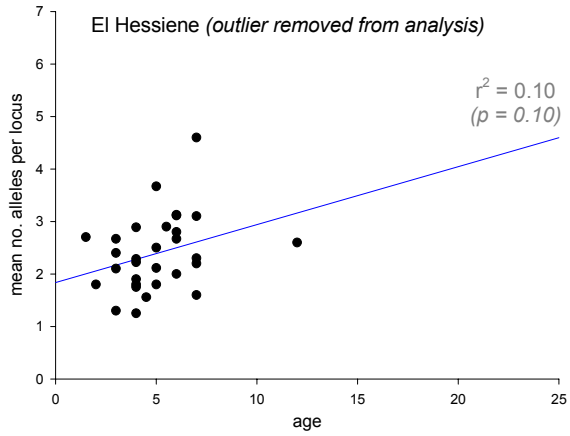
Figure 3.15. Correlation of cattle age and multiplicity of infection

The mean number of alleles per locus was plotted against the age of the host in months for isolates from the Tunisian village of El Hessiène **(i)** and the four districts in western Turkey **(ii)**. The linear regression line for each comparison is shown in blue. An outlying isolate was removed from the El Hessiène analysis, and the dataset reanalysed; no difference in the gradient of the linear regression line was observed. 'Goodness of fit' is indicated by r^2 values, which are generally low, indicating a high level of variance, although $p \leq 0.10$ in all populations except Aydın indicated statistical significance. When all five populations were combined a statistically significant correlation was also obtained ($n = 225$, $r^2 = 0.22$ and $p = 0.03$ (not shown)).

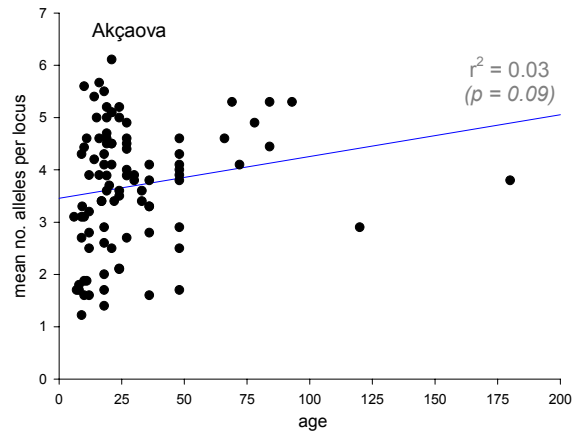
Figure 3.15. Correlation of cattle age and multiplicity of infection

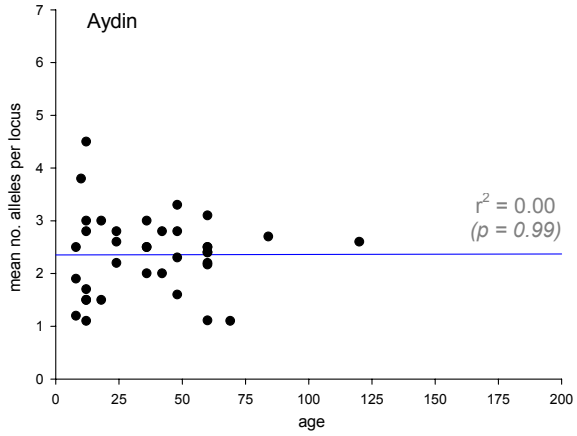
(i) Tunisia

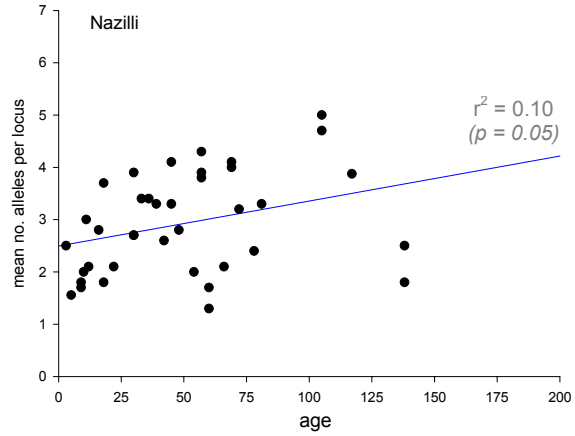




(ii) Turkey







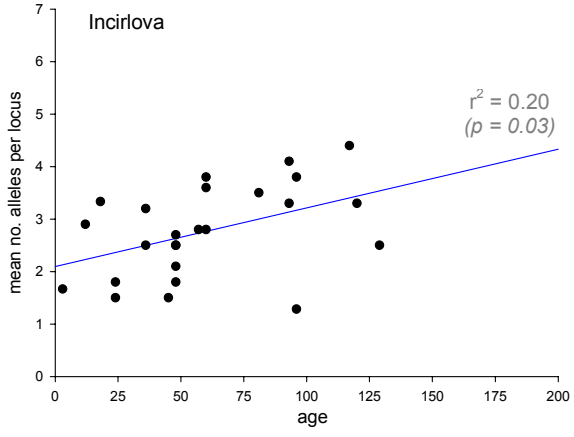
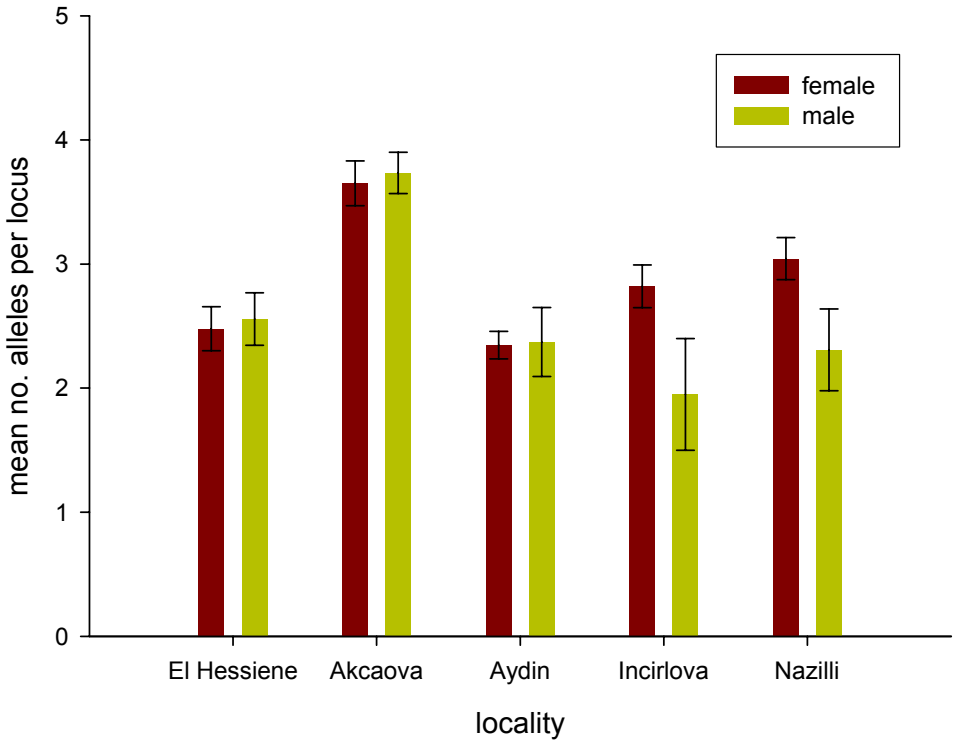


Figure 3.16. Multiplicity of infection with respect to locality and sex

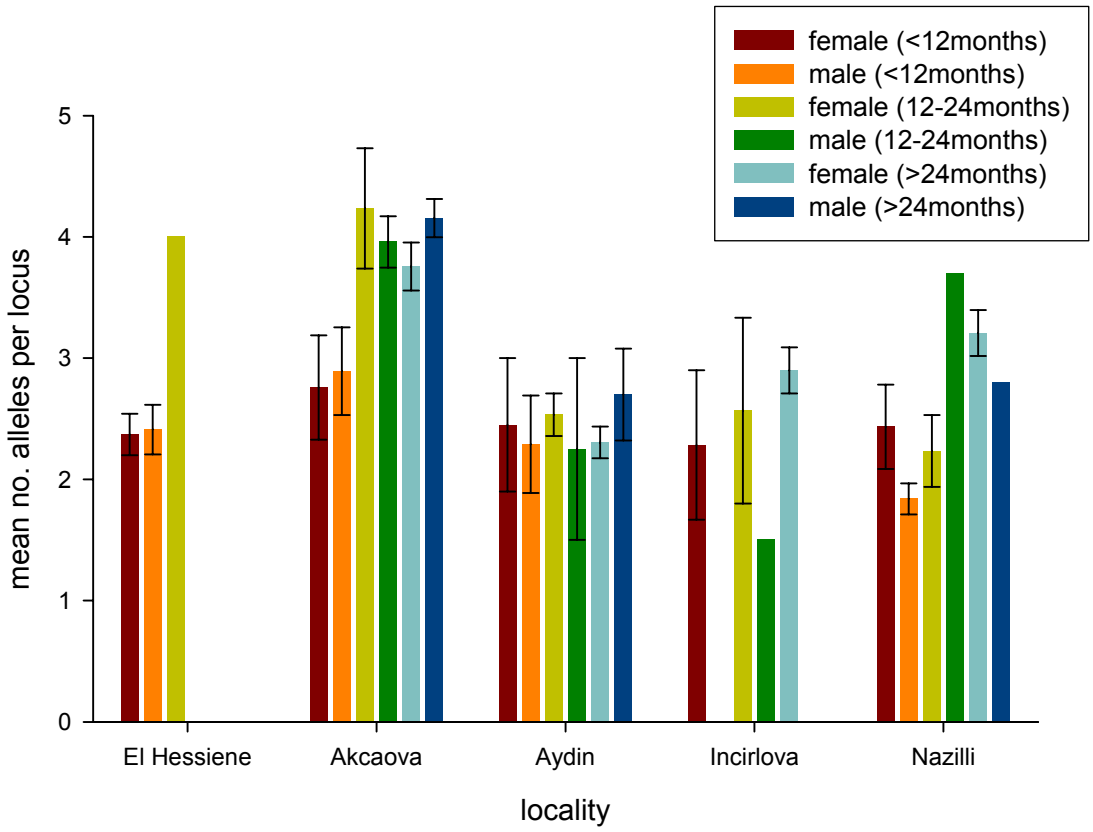
The multiplicity of infection was calculated for both male and female cattle in the village of El Hessiène in northern Tunisia and in the four districts in western Turkey. This was achieved by determining the mean number of alleles over the ten loci in each isolate and then calculating the mean value for each group. Each population was analysed as a whole **(i)** and then stratified by age and re-analysed **(ii)**. The standard error for each group is indicated; wide or absent error bars in the stratified analysis are largely due to small sample size.

Figure 3.16. Multiplicity of infection with respect to locality and sex

(i) Overall populations



(ii) Populations stratified by age



male population towards a lower multiplicity of infection. Although the multiplicity of infection is still higher in females than males less than one year, the sample size is too small to draw meaningful conclusions (males n = 4, females n = 3). The contradiction of the Incirlova and Nazilli results in Figure 3.16.(i) with that of the ANCOVA is attributed to using an increased dataset for this analysis, which encompasses more than the 199 isolates used in the ANCOVA analysis. This larger dataset was created from all the isolates where the host age was known and where data representing other host characteristics were not necessarily available. As previously indicated in Table 3.10., Figure 3.16.(i) illustrates that the multiplicity of infection was similar in several populations but was higher in Akçaova district.

It is possible that the increased multiplicity of infection observed in Akçaova is generated by the automated genotyping system, which did not record alleles that represented less than 32 % of the major allele (see Section 3.2.2.). Thus, if the samples from Akçaova rarely contained a predominant allele, many more alleles would have been recorded compared to samples with a predominant allele and a series of alleles less than 32 % of this allele. This would skew the estimate of multiplicity of infection. This potential effect of automated genotyping is illustrated in Figure 3.17. To investigate whether this was the case, the Genescan™ traces for marker TS5 were analysed in detail across the populations from the four regions. This marker was selected since alleles consistently differed in size by a full 6 bp motif, and therefore the software could confidently differentiate between each one. Additionally, TS5 showed a significantly greater multiplicity of infection in the Akçaova population compared to the other three districts and so reflected the differences in multiplicity observed when the mean number of alleles per locus was considered. The peak area, as a percentage of the predominant peak, for each allele from each sample in the four districts was calculated. The mean value for each ranked allele (i.e. predominant, second most abundant, third most abundant etc.) for each district was calculated and the data are presented in Figure 3.18. Little difference was observed among the four districts in Turkey. This indicated that the high level of multiplicity of infection in Akçaova was not attributed to an absence of a clearly predominant genotype as the peak areas of the secondary, tertiary etc alleles were a similar proportion of the predominant alleles in all districts. The results indicated that on average the peak relating to the secondary allele had an area of around 57 % of that of the predominant peak, thus the high level of multiplicity of infection in Akçaova was a real difference.

Figure 3.17. Illustration of a potential effect of automated genotyping

The automated genotyping protocol described in Section 3.2.2. may introduce errors when calculating the number of genotypes in an isolate in certain circumstances. A hypothetical example of this is illustrated opposite for a single locus.

(i) Several alleles present in approximately equivalent abundance

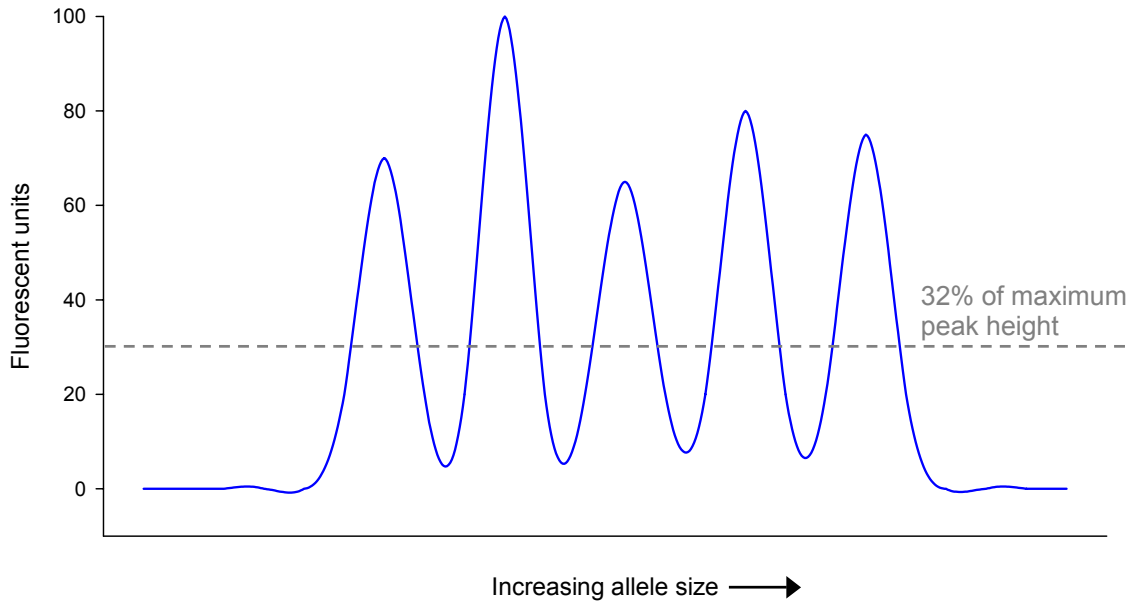
Five relatively abundant alleles are present at a locus, none of which is represented by a peak less than 32 % of the maximum peak height. Consequently five alleles are identified at this locus.

(ii) Single allele considerably more abundant than other alleles

The same five alleles are present in this trace, the second of which is far more abundant than any of the rest. Each of the four other peaks is less than 32 % of the maximum peak height and consequently they are removed from the analysis. Consequently, only a single allele is identified at this locus.

Figure 3.17. Illustration of a potential effect of automated genotyping

(i) Several alleles present in approximately equivalent abundance



(ii) Single allele considerably more abundant than other alleles

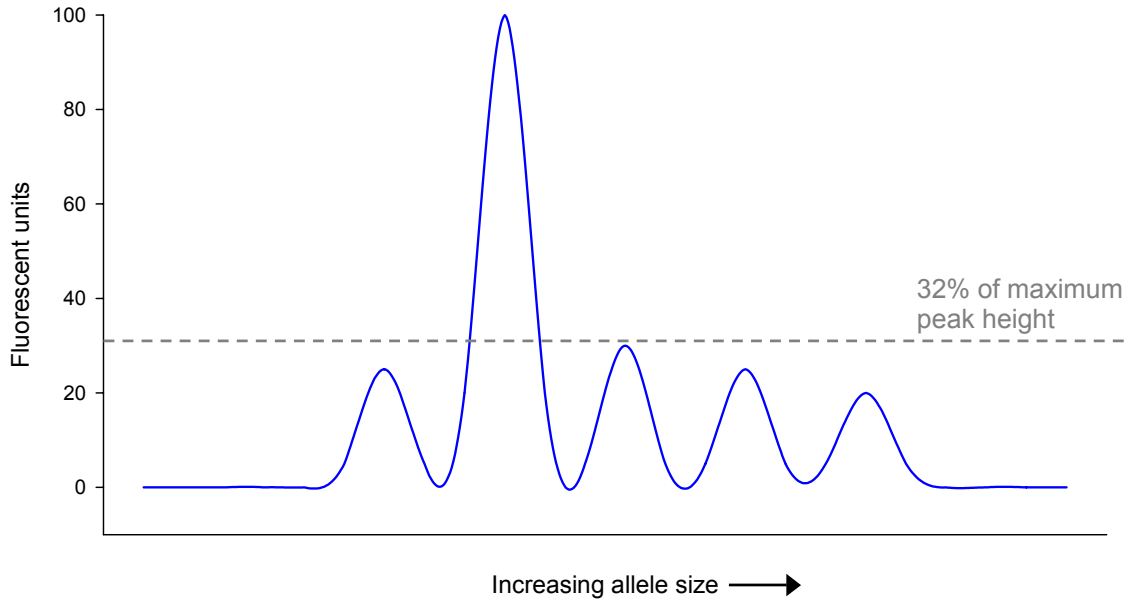
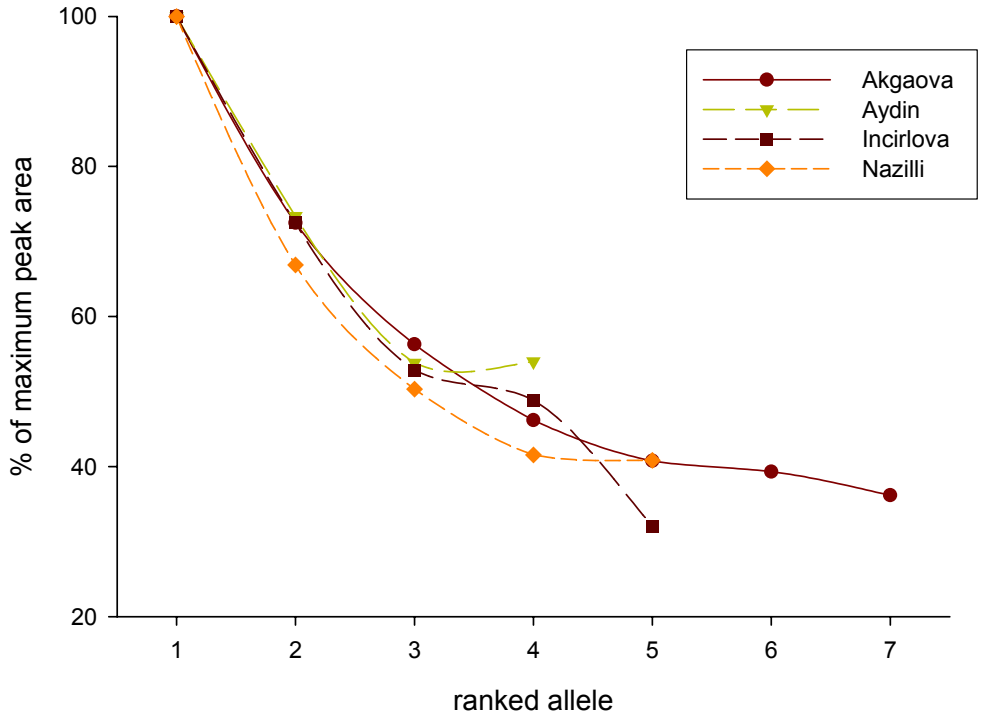


Figure 3.18. Relative proportions of TS5 alleles across Turkish districts

For every isolate in each district in Turkey, the allelic profile representing the TS5 locus was analysed. Alleles were ranked in order of descending peak area (i.e. predominant, second most abundant, third most abundant etc.) and the area of each peak was determined as a proportion of the maximum peak area. Little difference was observed between the four districts.

Figure 3.18. Relative proportions of TS5 alleles across Turkish districts



The co-variance analysis indicated that vaccinated cattle had a greater mixture of genotypes than unvaccinated cattle. Intuitively, it would be predicted that vaccination would lower the multiplicity of genotypes, although the introduction of the cell line vaccine may introduce the vaccine genotype into the population. To further investigate this phenomenon, two further analyses were performed. For each marker, the mean number of alleles at each locus was calculated for both vaccinated and unvaccinated cattle using (i) the entire Turkish dataset and (ii) isolates from Sariköy village in Akçaova. Rather than using the Akçaova district as a whole, the dataset representing Sariköy village was selected for analysis because isolates from that village contained a particularly high multiplicity of infection, averaging 4.10 alleles per locus (Table 3.10.) and these were isolated from approximately equivalent numbers of vaccinated (n = 28) and unvaccinated (n = 24) cattle. Furthermore, since isolates were collected from the same village, they were more likely to represent a sympatric population under the same level of challenge. The results of these comparisons are presented in Figure 3.19. In agreement with the co-variance analysis, an increased number of alleles was observed in vaccinated cattle at each locus when all Turkish isolates were analysed (Figure 3.19.(i)). However, when the village of Sariköy was studied in isolation, there was no statistical difference between the mean number of alleles per locus in vaccinated and unvaccinated cattle (Figure 3.19.(ii)). When reconciling these two analyses several points should be borne in mind – of the 40 isolates from vaccinated animals in the entire Turkish collection, the majority originated in Akçaova district and the multiplicity of infection in Akçaova was higher than in any other district (see Figure 3.17. and Table 3.10.). Hence, the co-variance result was misleading and arose because the bulk of the vaccinated cattle came from the district with the highest multiplicity of infection. This led to a false association between vaccination and multiplicity of infection.

3.3.6. Vaccine cell line genotyping

To investigate the impact of immunisation with Teylovac™ on field populations of *T. annulata* in Aydın province, a specific question was posed - can differences be observed between parasites isolated from vaccinated and unvaccinated cattle? Before this question was addressed it was useful to first appreciate the broad relationship between the Teylovac™ genotype and the Turkish field isolates. To this end, Teylovac™ and a series of seven cell line preparations from Turkey and several other countries were analysed and their multilocus genotypes determined (Table 3.13.). This included a number of cell lines developed for vaccination in Tunisia but never deployed in the field. Seven of the eight cell lines displayed a single allele at each locus, notwithstanding a failure to amplify

Figure 3.19. Multiplicity of infection in vaccinated and unvaccinated Turkish cattle

The multiplicity of infection (i.e. the mean number of alleles per locus) was calculated for vaccinated and unvaccinated cattle for each marker in two overlapping collections of isolates. The maximum (Max), minimum (Min) and mean number of alleles per locus was also determined for each collection.

(i) All Turkish isolates

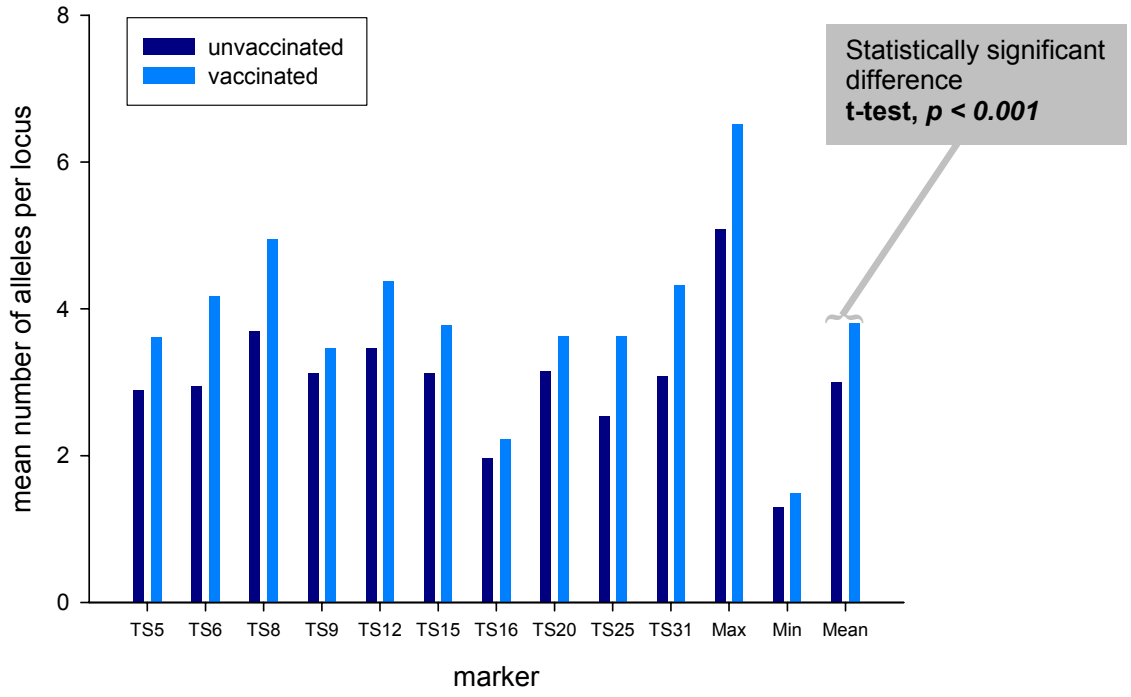
An increased multiplicity of infection in vaccinated cattle was indicated over all ten loci when analysing the entire Turkish isolate collection. A significant statistical difference in the mean value between groups was demonstrated using Student's t-test ($p < 0.001$).

(ii) Isolates from Sariköy village

The dataset representing Sariköy village was selected for independent analysis since it represented approximately equivalent numbers of vaccinated (n = 28) and unvaccinated (n = 24) cattle sampled from a single site. The multiplicity of infection was similar in vaccinated and unvaccinated cattle over individual loci and no statistical difference between the mean number of alleles per locus in each group was detected using the Mann-Whitney test ($p = 0.927$).

Figure 3.19. Multiplicity of infection in vaccinated and unvaccinated Turkish cattle

(i) All Turkish isolates



(ii) Isolates from Sariköy village

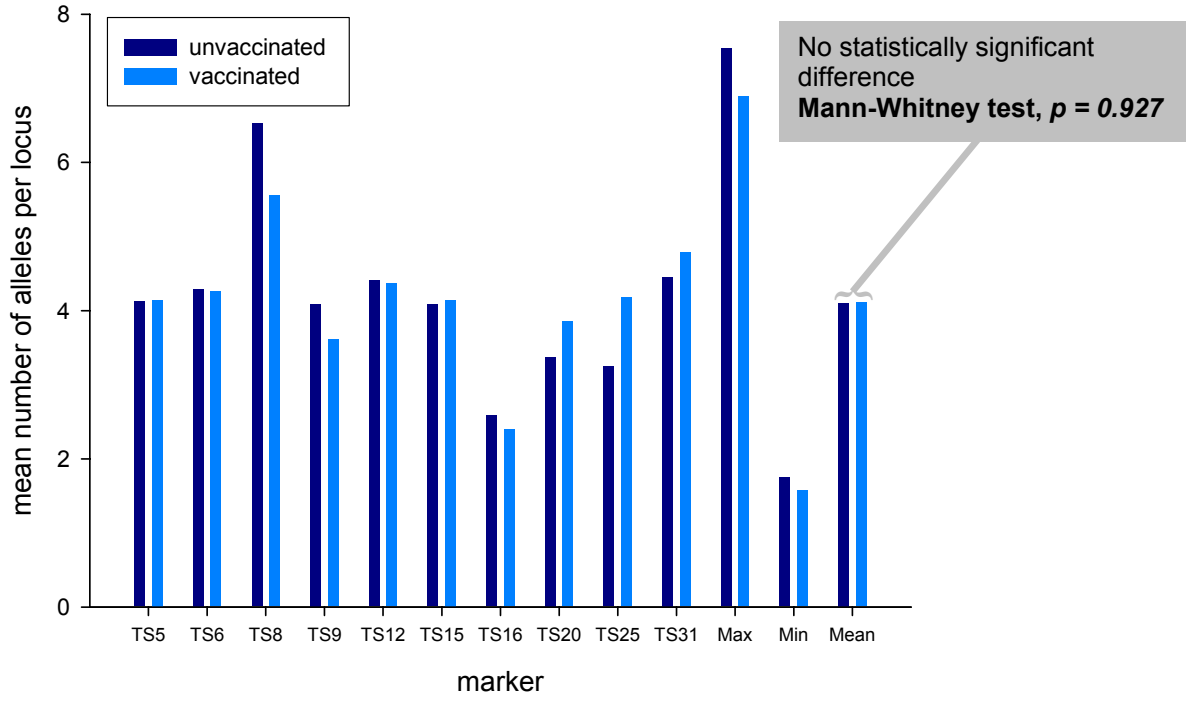


Table 3.13. Genotyping of cell-lines developed for vaccination

DNA preparations from eight cell lines developed as potential vaccine lines against tropical theileriosis were genotyped at ten micro- and mini-satellite loci. Minor alleles present at three loci in the Béja isolate are denoted in parenthesis. The three Tunisian cell lines have not been used for immunisation in the field, however the other five cell lines have been deployed in their country of origin. This includes Teylovac™, a commercial Turkish vaccine reportedly based on the Pendik cell line. Teylovac™, highlighted in blue, is virtually identical to the high passage laboratory culture of the Pendik cell line.

Table 3.13. Genotyping of cell-lines developed for vaccination

Country	Cell line	Passage	Marker / alleles (bp)									
			TS5	TS6	TS8	TS9	TS12	TS15	TS16	TS20	TS25	TS31
Turkey	Diyarbakir	p107	270	436	237	352	242	308	353	225	250	257
"	Pendik	p313	264	389	305	352	282	356	350	236	218	286
"	Teylovac	unknown	264	388	305	352	282	356	349	236	218	285
Tunisia	Batan2 *	p198	276	NA	324	354	237	164	349	262	241	257
"	Béja *	unknown	282 (270)	386	237 (261)	372	253	262	351	209 (229)	213	333
"	Jedeida4 *	p200	270	360	298	373	254	284	NA	238	241	265
India	Ode	p58	252	≈600	270	354	336	238	361	222	209	318
Spain	Caceres	p107	288	NA	270	357	315	238	349	310	NA	265

NA = no amplification

* developed but never deployed

several loci when Caceres, Batan2 and Jedeida4 were used as a template. The Béja cell line exhibited a secondary allele at loci TS5, TS8 and TS20 indicating that two or more parasite genotypes were present. It can be seen that the commercial vaccine preparation, Teylovac™ had a very similar profile to that of the Pendik cell line, with allelic sizes at seven loci being identical and the remaining three loci showing a difference of only 1 bp. Pendik and Teylovac™ preparations were therefore considered to be genotypically identical with minor artefactual differences in MLGs due to mis-sizing alleles. This was attributed to slight variation in allele binning between the automated system, used to genotype Teylovac™ and the manual system, previously used to genotype the Pendik cell line. To illustrate the relationship of cell lines developed for immunisation within Tunisia and Turkey to the genotypes of the field isolates from each country, PCA analysis was undertaken. PCA analysis had already differentiated Tunisian and Turkish parasite isolates (Figure 3.8.), therefore it was considered appropriate for revealing any underlying relationship between the cell lines and these field populations. The results of this analysis are presented in Figure 3.20. The Pendik and Teylovac™ MLGs were found within the body of the Turkish samples in the left quadrants, while the Béja genotype lay towards the opposite pole on the first axis, within the Tunisian cluster. Batan2, Jedeida4 and Diyarbakir genotypes were found at the interface between Tunisian and Turkish genotypes. When the two genotypes from Spain and India were incorporated into the analysis, they located at this interface region (data not shown); this may be interpreted to indicate – foreign genotypes unrelated to either population fail to cluster within either of the countries. The results of the PCA analysis suggested that the Béja cell line was similar to the Tunisian isolates and that the Pendik / Teylovac™ cell line was similar to those from Turkey. Therefore it was concluded that each of these cell lines was related to the country in which it originated.

To further investigate the relationship of Teylovac™ to field isolates, the entire Turkish population was grouped into isolates from vaccinated and unvaccinated individuals and the frequency of Teylovac™ alleles in each group was determined. The results of this analysis are presented in Figure 3.21.(i). Alleles of loci TS6 and TS15 from Teylovac™ were not represented in either group. With the exception of TS8 and TS20, for each of the six loci where a Teylovac™ allele was present, it was at a higher frequency in vaccinated compared with unvaccinated animals. However, at five of the six loci, no statistical significance was revealed between vaccinated and unvaccinated groups using the Chi-squared test (Figure 3.21.(i)). Therefore, it can be concluded that the immunising genotype was not over-represented in the vaccinated cattle over the entire Turkish population.

Figure 3.20. Relationship of cell lines to field isolates

PCA analysis was repeated on the multi-locus genotype data representing the new Tunisian and Turkish populations also incorporating the six multi-locus genotypes corresponding to Tunisian and Turkish attenuated cell lines. The two principal axes generated by this analysis are presented opposite, demonstrating the distribution of these cell lines in relation to the field isolates. Teylovac™ and Pendik cell lines were considered identical, with different but proximal locations attributed to slightly variant multi-locus genotype data.

Figure 3.20. Relationship of cell lines to field isolates

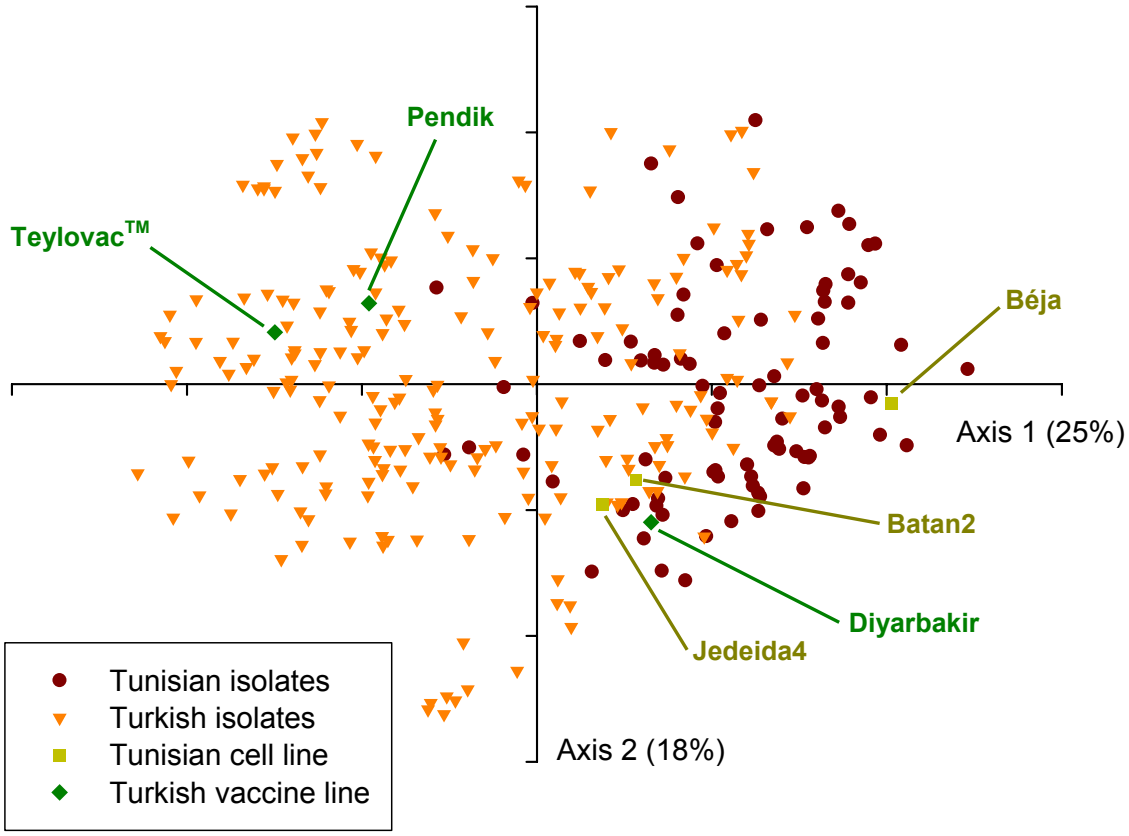


Figure 3.21. Frequency of Teylovac™ alleles in field population

The frequency of the alleles corresponding to the Teylovac™ cell line were determined in parasite populations isolated from vaccinated and unvaccinated cattle. Two collections of field isolates were used –

(i) Turkey

This group comprised the entire collection of isolates from Turkey. A Chi-squared test was performed on the eight loci where the immunising allele was present in the population. At loci where both allele frequencies were low, the test had limited power to detect differences. Only for locus TS12 was a statistically significant difference (*p = 0.011*) revealed between the unvaccinated and vaccinated populations, i.e. the allele was more frequent in the vaccinated group.

TS5: $\chi^2_1 = 0.739$, *p = 0.390*
TS8: $\chi^2_1 = 0.254$, *p = 0.614*
TS9: $\chi^2_1 = 1.137$, *p = 0.286*
TS12: $\chi^2_1 = 6.511$, *p = 0.011*
TS16: $\chi^2_1 = 2.919$, *p = 0.088*
TS20: χ^2_1 = *0.0822, p = 0.774*
TS25: $\chi^2_1 = 0.644$, *p = 0.422*
TS31: $\chi^2_1 = 0.000757$, *p = 0.978*

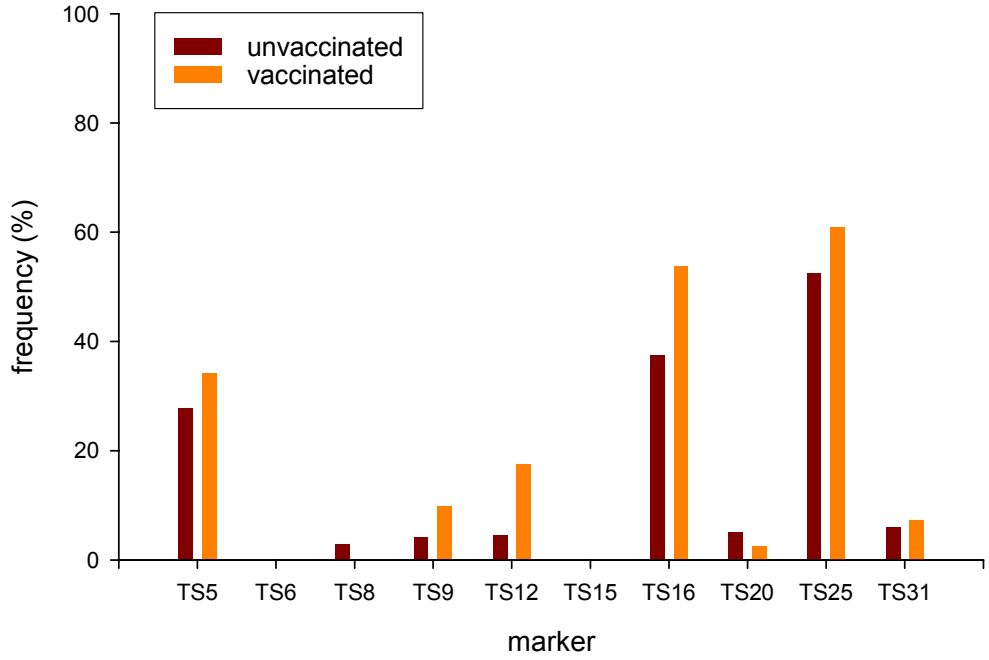
(ii) Sariköy village

The dataset representing Sariköy village was selected for independent analysis since it represented approximately equivalent numbers of vaccinated (n = 28) and unvaccinated (n = 24) cattle sampled from a single site. Due to the smaller sample size, Fisher's exact test was performed on the seven loci where the immunising allele was present in the population. No significant differences in allele frequency were detected between the unvaccinated and vaccinated groups.

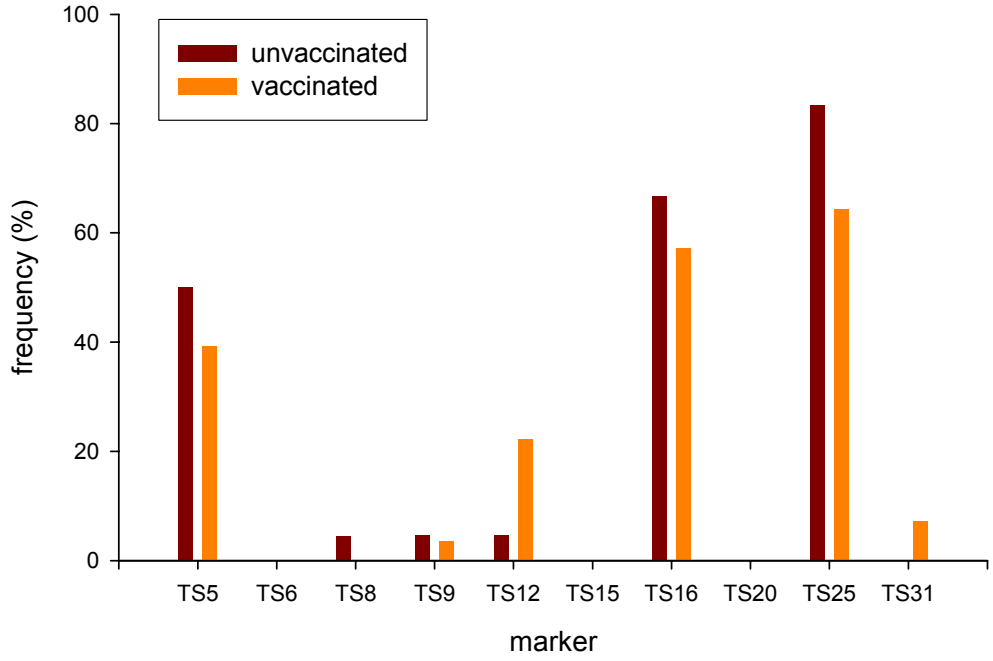
TS5: *p = 0.577*
TS8: *p = 0.460*
TS9*: p = 1.000*
TS12: *p = 0.112*
TS16: *p = 0.573*
TS25: *p = 0.209*
TS31: *p = 0.497*

Figure 3.21. Frequency of Teylovac™ alleles in field population

(i) Turkey



(ii) Sariköy village



However, it can be seen from Table 3.11.(ii) that only around 20 % of the Turkish cattle population was vaccinated. The proportion of vaccinated cattle also varied from district to district and between villages. Therefore, similar to the study on multiplicity of infection, the isolates from Sariköy village in Akçaova were selected for independent analysis. With approximately equal numbers of vaccinated (n = 28) and unvaccinated (n = 24) cattle sampled from this single site, it represented an ideal unbiased sample collection. The results of the analysis using the isolates from this village are presented in Figure 3.21.(ii). In general agreement with the results from the entire Turkish population, alleles of loci TS5, TS16 and TS25 were relatively frequent in the population, however, frequencies were slightly higher in the unvaccinated group. Fisher's exact test was used to determine whether this difference in frequency between the vaccinated and unvaccinated groups was significant; for all loci, no significant difference was revealed. Finally, to investigate the relationship between MLGs of isolates from vaccinated and unvaccinated cattle, the PCA representing MLGs of Turkish isolates was relabelled to indicate the vaccine status of the host and the data are presented in Figure 3.22. The lack of a discrete clustering of the isolates from vaccinated animals supported the conclusion that vaccination had little impact on genotypic diversity in the field.

3.4. Discussion

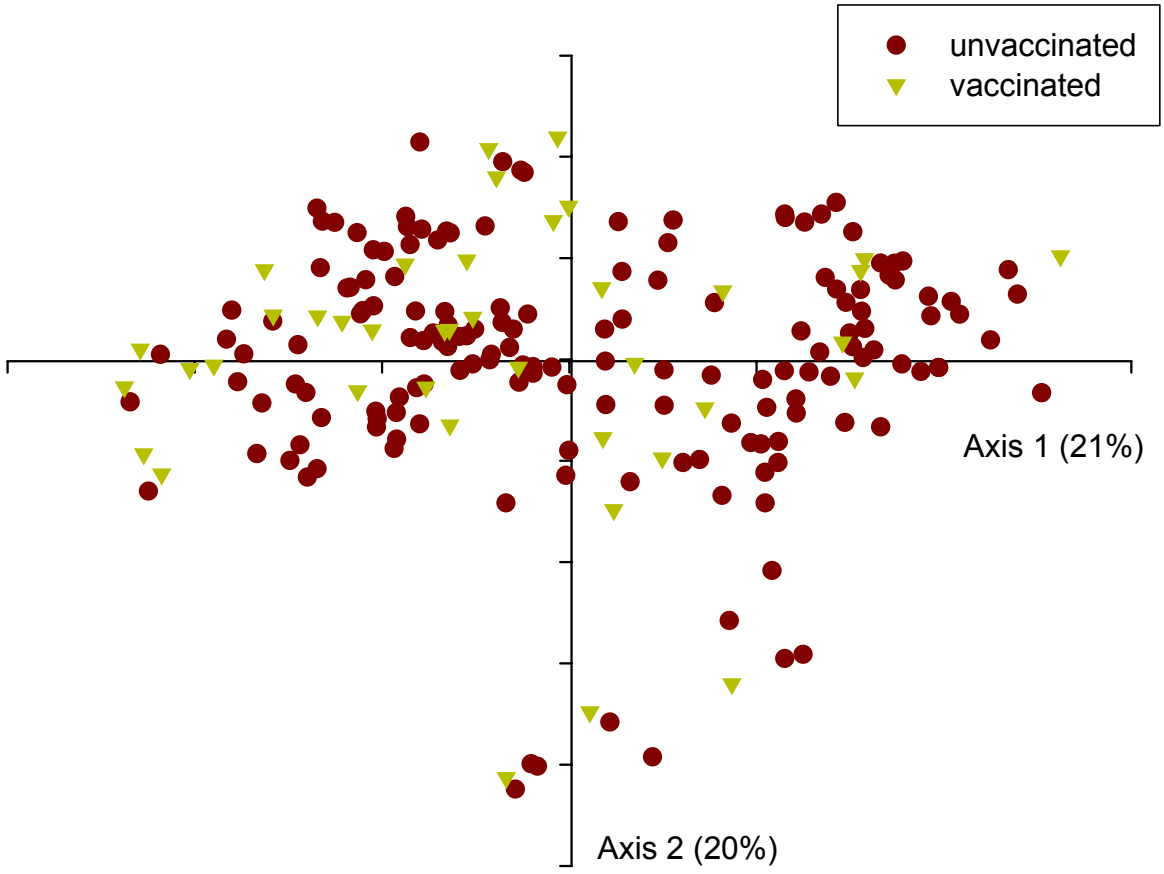
3.4.1. Comparison with initial population genetic study

The principal conclusion arising from the preliminary study, i.e. that genetic differentiation was detectable between geographically separated populations of *T. annulata*, was confirmed using this larger, more extensive dataset. A moderate level of differentiation was indicated in both studies (F_{ST} = 0.049 / 0.052), which decreased when each country was examined independently. The initial study, which encompassed three diverse sampling areas in Tunisia estimated F_{ST} at 0.023, while the new analysis estimated F_{ST} at 0.017 between the three contiguous farms at El Hessiène. These results may be interpreted together as demonstrating a gradation of genetic differentiation, ranging from a low level found in one sampling site (El Hessiène) to an intermediate level when sites from different areas in the same country are analysed (East, West and Central Tunisia) to a higher level when samples from different countries are compared. This may be taken as evidence of a degree of genetic isolation between countries consistent with geographical and possibly trade barriers hindering the free flow of genetic material. In particular, this new study provided an improved account of diversity of *T. annulata* populations within Turkey, due to the much larger, structured sampling regime. Genetic differentiation, which was not

Figure 3.22. PCA of vaccinated and unvaccinated Turkish cattle

Principal co-ordinate analysis was repeated on the multi-locus genotype dataset representing the new sample collection from Turkey. The two principal axes generated by this analysis are presented opposite with data points colour-coded to represent isolates from either vaccinated or unvaccinated cattle.

Figure 3.22. PCA of vaccinated and unvaccinated Turkish cattle



previously detected in the initial 13 Turkish isolates, was found when four neighbouring districts were analysed in detail. The resultant F_{ST} value of 0.028 was comparable to the level of genetic differentiation detected within Tunisia. The new results underscored a caveat of the initial study, i.e. caution should be exercised when interpreting results when based on a limited sample size.

Consistent with the first study, significant linkage disequilibrium was detected when Tunisian and Turkish isolates were pooled. This was reduced when the Tunisian collection was analysed on a site-by-site basis, however, in contrast to the previous study, a degree of linkage disequilibrium was still observable across Tunisia as a whole. The major contrast with the initial study was the level of linkage disequilibrium detected both across the entire Turkish population and within three of the four districts sampled in Turkey. I^S_A values indicated that LD was present within districts at a higher level than over the entire population. It is worthwhile considering some of the various reasons that may explain LD, such as a low level of genetic exchange, population sub-structuring, recent immigration, selection and inbreeding (including self-fertilisation). However, the first and most obvious was physical linkage between loci. The ten marker loci used in this study were distributed over the four chromosomes. Two markers, TS8 and TS9 were originally identified on the same large contig during assembly of the genome and being separated by only 38 kb, these were by far the two closest loci. It was demonstrated that genetic linkage between these loci did not influence the analysis by comparing the linkage results using all ten loci to results with TS9 marker data removed. Although slight variation was observed between the values, which represented both minor increases and decreases in I^S_A over the different population comparisons, it was concluded that LD was an intrinsic feature of the populations where it was identified. Genetic differentiation and geographical sub-structuring between Tunisian and Turkish populations was consistent with LD being detected at this level. Significantly, both LD and LE were observed in different areas of Tunisia and Turkey. In Tunisia, populations from two farms at El Hessiène and from Béja each displayed LE when analysed independently, with LD indicated only in the Hassine farm. In Turkey isolates from the district of Incirlova and the village of Osmanbuku in Aydın were also in LE, in contrast with the rest of the populations. In these regions it can be inferred there was no impediment to gene flow and a truly panmictic population structure was indicated. In the areas where LD was observed, there was clearly some form of restraint on free associations of alleles within each district. With LD exhibited in three of the four Turkish districts and population differentiation demonstrated (Table 3.4.), it can be immediately inferred that there was a degree of restriction to gene flow between

districts. However, it must be emphasised that I^S_A values were relatively small in all the populations in LD and therefore a very high background level of genetic exchange existed with genetic linkage being limited. Clearly, the parasite population did not exhibit a clonal structure. Moreover, the stratified, subtractive linkage analysis (Table 3.8.) indicated that in each country, a limited number of clonal or highly similar genotypes were not responsible for causing LD. Consequently, a general epidemic-type population structure was also ruled out. It was therefore reasonable to conclude that the underlying population structure for *T. annulata* was panmixia and this view is discussed more broadly in Section 6.3. Panmixia was immediately evident in areas in LE and it was only slightly disturbed in the various areas with limited LD. So how should the LD exhibited in many of the localities be interpreted? It is possible that when isolates from within a single district are studied they represent more than just a single interbreeding population. This may be due to micro-geographical sub-structuring whereby herds of cattle are kept in isolation and are infected by populations of ticks, which do not move between areas. This would not explain the situation in Hassine farm in Tunisia, for example, where a genuine cohort of cattle was used to isolate parasite material. Additionally, the high level of heterozygosity, high multiplicity of infection and extreme diversity observed in each population would limit the effect of random genetic drift and therefore this is unlikely to explain LD. An alternative explanation is admixture of populations, i.e. the Wahlund effect. This would occur when a distinct parasite population was recently introduced into an area and would be likely due to the introduction of infected cattle from other localities. It would take a period of time before the pre-existing genotypes and 'foreign' genotypes recombined to homogenise the local population and so restore LE. This is in part due to the ecology of the vector and parasite, where only a single generation of new parasites are produced per year. Tick larvae engorge on cattle in the autumn, acquiring the parasite following which, sexual recombination occurs resulting in a new generation of *T. annulata* being inoculated into cattle the following disease season. Therefore, only a single round of reassortment of alleles occurs per year. The tick vector is unlikely to travel far in the course of a year, unless major cattle movements occur. Therefore, local gene pools probably undergo only minor changes over the course of a year, with the most significant effect being produced by influx of new cattle harbouring new genotypes of *T. annulata*. This may be one explanation for the distinct cluster of isolates from the village of Sümer Mah in the Turkish district of Nazilli. As well as being relatively distinct from the majority of the Turkish isolates (Figure 3.9.(ii)), they were shown to be highly related to each other, with a low number of alleles identified over the ten loci (Table 3.5.). Additionally, these isolates exhibited a low multiplicity of infection (Table 3.10.) and as

discussed earlier, this may represent an epidemic population structure within this locality. An epidemic population structure on this scale is perfectly compatible with novel genotypes, perhaps recently introduced to a naïve cattle population.

It is also worth considering whether a selective pressure may have influenced population diversity in each district. For example, it is theoretically possible that a selective sweep resulting from vaccination or drug usage resulted in selection of particular genotypes and that they were over-represented in the population, thus elevating LD. Although possible, the impact of vaccination could not be detected in this study when isolates from immunised and unimmunised cattle were compared (Figures 3.19., 3.21. and 3.22.). Moreover, one would suspect that the effect of selection would have been detected in the stratified linkage analysis (Table 3.8.). Thus, a single highly related cluster from the Turkish dendrogram (Figure 3.12.) might have been predicted to represent descendents of a selected genotype; however, no such cluster was evident. Furthermore, for LD to be maintained in the population, genes with alleles subject to selection would have to be physically linked to two or more marker loci.

Perhaps the most likely explanation for LD was that a degree of inbreeding was occurring. A considerable number of Turkish isolates were shown to share a large proportion of alleles with other isolates from the same district (Table 3.7. and Figure 3.10.). This was a remarkable degree of similarity when the global polymorphism of each of the ten markers was considered. These results demonstrated a gradation of similarity and suggest mixtures of genotypes exhibiting varying levels of kinship. This may simply represent related individuals recombining, or it may be associated with a more extreme form of inbreeding, self-fertilisation. The fact that only two isolates in the population share a common MLG suggests that self-fertilisation, if present, is masked by recombination in ensuing generations. In *Plasmodium*, the local transmission intensity is believed to influence linkage and in areas of low transmission, increased LD is observed (Anderson *et al.* 2000a). However, in this study, low multiplicity of infection in a district (Table 3.10.), which may indicate low transmission intensity, did not correlate with increased LD (data not shown). However, a structured field study involving extensive epidemiological surveying would be required to properly distinguish between areas of high and low transmission intensity. It is also important to consider that related parasites may simply have been sampled. For example, if an animal was infected with, say two parasite genotypes and a large number of ticks fed on the animal, once the ticks became infected, recombination between these two parental genotypes in different ticks would have

generated a large number of siblings. Therefore in the following disease season several animals would be inoculated with sibling *T. annulata* genotypes and therefore the parasite population would show a departure from LE. Clearly, a structured epidemiological study is required to assess the impact of these factors on parasite populations in the field.

In light of results presented in the previous chapter (Table 2.8. and Figure 2.11.), it is possible that similarity and LD may have been underestimated in this study considering each isolate represented a heterogeneous parasite population. It was shown in Chapter Two that a ten marker MLG, derived from an Ankara piroplasm DNA extract, shared only five and four of its principal alleles with its two derived clones. Additionally, homologous piroplasm and cell line extracts were shown to share as few as five alleles, although they contain a proportion of identical genotypes. Therefore it could be argued that homologous preparations, which likely share some genotypes may potentially only exhibit 50 % identity. With complex mixtures of genotypes present in every one of these new blood samples, one must consider that when a pair of isolates match at for example, five loci, the mixtures of genotypes in each of the isolates may be very highly related indeed. In future, it would be interesting to examine the complete allelic profile of pair-wise combinations of isolates, whose predominant MLGs exhibit a high level of similarity. This would reveal whether the predominant alleles at non-matching loci were present in the other isolate as secondary components. However, in order to definitively identify the constituent haploid genotypes in a mixed infection, it would be necessary to generate clones from the isolate and determine each of their genotypes individually.

In the initial study, cluster analysis of the predominant MLG failed to group isolates based on geography. Evidence for geographical sub-structuring was found when the full allelic profile of each isolate was used for the similarity analysis. Using only the predominant allele at each locus, the new dataset was used to demonstrate geographical sub-structuring between countries (Figure 3.8.) and sub-structuring among Turkish districts (Figure 3.12.). This was attributed to both the increased sample size, especially in Turkey, and the structured nature of the sampling, with multiple, related isolates being collected from the same site. Similar to the initial study, complete genetic isolation between Tunisia and Turkey was not observed (Figure 3.8.). Some Tunisian isolates were found within the Turkish cluster and an interface region was present, the significance of which is difficult to assess. This region represented isolates that possessed a limited number of alleles, which occurred at a high frequency commonly found in the other country. Genotypes representing two diverse cell lines from Spain and India were also found to cluster in this

interface region. Therefore, rather than evidence for a degree of direct gene flow between Tunisia and Turkey, the interface region may simply reflect genotypes that have a limited identity to either of the two populations which have been analysed in this study. As the disease is present in countries that lie between Tunisia and Turkey (Libya, Israel, Iran and Iraq), it is possible that while there is no direct cattle movement between Tunisia and Turkey, there is movement between neighbouring countries allowing a limited level of parasite movement across the whole region. This in turn would allow a limited degree of gene flow of *T. annulata* between the two countries. Alternatively, the lack of clear differentiation between the countries may reflect ancestral polymorphisms in the parasite population or an interchange of genetic material in the distant past. However, it would be unwise to over-interpret the incomplete separation of each population, and it must be emphasised that the significant conclusion from the analysis is that macro-geographical sub-structuring has been confirmed.

The results presented in this chapter represented a completely novel dataset and although the same markers were employed as in Chapter Two, a new method of identifying and scoring alleles was implemented. It is important to compare the results of both studies with reference to each individual marker's ability to discriminate between populations. Even although a large number of novel alleles were identified in the more extensive study, it can been seen in Figure 3.6.(i). that a similar level of heterozygosity was indicated in both studies across all markers. Population differentiation, as indicated by G_{ST}', was of a similar level across several markers with the highest level indicated in both studies with marker TS25. For this locus, the different allele frequencies within each country were remarkably similar over both studies (Figure 2.7. and Figure 3.5.). In both datasets, the 213 bp allele was the major allele in approximately 45 % of Tunisian isolates, with the alleles of 241 bp and 250 bp representing the major alleles in 40 % of samples. In contrast, the 218 bp allele was by far the most frequent in the Turkish population over both studies. Markers, such as TS25, with few alleles that are present at a relatively high frequency are considered optimal for population genetic studies. In the analysis of linkage disequilibrium in populations, markers with a large number of low frequency alleles lead to predictions of even lower frequencies of allele combinations when two such loci are analysed. If the sample size is insufficient, then such combinations of alleles are not predicted to occur and therefore it is concluded that the population is in LE, when in fact this hypothesis has not been properly tested. In contrast, loci with several alleles at relatively high frequency have testable predicted frequencies of allelic combinations and so give a more robust analysis of LE. Therefore, perhaps, this particular marker is an accurate

reflection of the population differentiation between the countries. However, these results clearly demonstrate the capability of the new automated genotyping system to consistently and accurately identify alleles defined in the preliminary study.

The field study comprised parasites isolated from both carrier animals and clinical cases of tropical theileriosis. Unfortunately, it was not known which isolates came from each class. Although this is clearly a shortcoming of the dataset, on balance it probably made little difference to the genetic analysis. This is because the *T. annulata* populations in each country were shown to exhibit extreme diversity and to possess a primarily panmictic structure. Therefore, genotypes were considered to be transient, with almost every individual isolate associated with a novel MLG. Furthermore, with the micro- and mini-satellites considered as selectively neutral, there was no reason to suspect that 'carrier' MLGs would be distinguishable from 'virulent' MLGs. For such an association to exist, the locus of a gene with alleles corresponding to either hypothetical phenotype would have to be genetically linked to one or more of the marker loci. In the unlikely event of this being the case, the MLG data would be biased solely at the linked loci. Furthermore, unless linkage was very strong, linkage disequilibrium would quickly decay in a randomly mating, highly diverse population. However, it is easier to envisage the influence of either carrier or clinical case status affecting parasite dynamics and apparent multiplicity of infection within a host. Hypothetically, if clinical disease involves a single predominant genotype clonally expanding or differentiating faster than other genotypes, it would be over-represented in the red cell population (analogous to Figure 3.17.(ii)). Hence, acutely affected clinical cases may display a reduced multiplicity of infection compared to carrier animals. Conversely, it may be argued that carrier animals would under-represent the number of mixed genotypes, if the number of parasite genotypes were selectively reduced over time. Unfortunately neither hypothetical situation could be studied using this dataset. However, various other factors, such as host age and vaccination status were investigated to explain the differing levels of multiplicity of infection observed across the collection of isolates from both countries.

3.4.2. Host age, vaccine status and multiplicity of infection

In this study multiplicity of infection was shown to differ between the different sampling sites. In Tunisia a similar rate was observed over all four sampling sites while in Turkey variation between villages was marked and in general, multiplicity of infection was shown to be greater. Perhaps the most likely explanation for this variation was differential transmission intensity between sampling sites. Several factors may determine transmission

intensity in a particular locality including tick prevalence, incidence of tick infection and cattle management practices. To an extent, transmission intensity may be related to differential states of endemicity, as discussed in Section 1.7. It could be argued that the lack of increase in multiplicity of infection with age in Aydın represents high challenge and endemicity. However, the mean level of infection in Aydın is the lowest of the four districts, casting doubt on this theory. It would be interesting to determine whether sites of high multiplicity of infection, such as Sariköy village, exhibited endemic stability for tropical theileriosis. In contrast, a village such as Sümer Mah, with less than half the multiplicity level in Sariköy may be predicted to be endemically unstable. That fact that the Sümer Mah isolates were highly related to each other suggests that in addition to low multiplicity of infection, a predominant genotype was sweeping this area. Taken together, these observations may in fact indicate an epidemic population structure existing at a micro-geographical scale. This would be consistent with a naïve population facing a sudden challenge, however one may only speculate about the underlying reason. For example, an epidemic may involve influx of infected ticks or cattle into a previously sterile area, movement of naïve cattle to infested pasture or simply a sudden breakdown of control measures such as vaccination or acaricide treatment. These findings therefore highlight the need for future studies to take account of vector ecology and agricultural management practises. One may then be able to measure the level of challenge and test whether a correlation exists between this parameter and multiplicity of infection.

The analysis of co-variance of the large dataset in Turkey, incorporating a range of host variables, suggested that age and vaccination status of the host might explain the multiplicity of *T. annulata* infection (Table 3.12. and Figure 3.14.). Analysis of individual districts in Turkey and the farms at El Hessiène in Tunisia supported the association with host age (Figure 3.15.) but discounted the influence of host vaccination status. The Tunisian sample demonstrated the build-up of infection over the course of a single disease season, while the Turkish findings were consistent with the view that in endemic areas, individual animals are exposed to life-long re-challenge. It would have been interesting to examine if a plateau effect was observed with increasing age. Such a plateau may be predicted where (1) a saturation point is reached when an individual is exposed to virtually all the diversity of the local area and (2) when eventually enough cross-protective immunity builds up to prevent further novel genotypes infecting. Unfortunately, this study had little power to detect such a phenomenon due to the variance in multiplicity of infection between individuals and the fact that few cattle older than six years were sampled.

The numbers of indigenous cattle were too small to draw conclusions about how breed may affect multiplicity of infection. In addition, no clear evidence was obtained to suggest that the sex of the host influenced the number of harboured genotypes. Although the co-variance analysis suggested that males contained more genotypes than females across the entire dataset, when individual areas were analysed, no difference was observed in El Hessiène (Tunisia), Akçaova and Aydın (Figure 3.16.(i)). In complete contrast, in the districts of Incirlova and Nazilli females showed evidence of a higher multiplicity of infection compared to males, although considerable variance in values for males was indicated by the larger error bars. One would predict that differences in parasite multiplicity between male and female cattle would be related to differential management practices rather than an underlying biological variation between the sexes. For example, dairy cattle may be taken out to graze on a daily basis and are likely to be moved between different pastures while on the other hand a stock bull, is more likely to be kept primarily in a secure paddock or enclosure. Unfortunately, data relating to the management practices of the farms from which samples were obtained was unavailable.

The association between positive vaccination status and a multiplicity of infection was also shown to be confounded by other variables. When the multiplicity of infection in the entire Turkish sample was related to vaccine status (Figure 3.19.), a higher mean number of alleles were indicated across all loci. However, as can be seen in Figure 3.16.(i) and Table 3.10., the isolates from the Akçaova population (where the majority of vaccinated cattle were located) generally possessed a high number of genotypes irrespective of their vaccination status. To control for this factor, a roughly equivalent number of unvaccinated and vaccinated cattle from the village of Sariköy in Aydın were analysed. No statistical difference between the two groups was demonstrated raising a new question- is there an association between the use of cell line vaccination in a district and an increased multiplicity of infection in that district as a whole? It is important to remember that the mean number of alleles per locus per individual only provided an index to the multiplicity of infection and that an upper limit of twelve alleles was imposed on each marker. Therefore, it would be a complete misinterpretation to assume that the isolates in Akçaova simply had one extra genotype per individual and that this corresponded to the immunising cell line. Furthermore, comparison of the Teylovac™ cell line with Turkish field isolates indicated that the immunising genotype did not establish at a detectable level in the field population and suggests that it does not induce a carrier state.

3.4.3. Vaccine cell lines

Similarity and PCA analysis demonstrated that the Béja and Teylovac™ cell lines genotypes clustered within Tunisian and Turkish populations respectively (Figure 3.21.). Correspondingly, across nine loci, the Béja cell line alleles were present at a markedly higher frequency in Tunisia than in Turkey and conversely, at seven loci the Teylovac™ alleles were present at a slightly higher frequency in Turkey than in Tunisia (data not shown). These findings may simply be explained as the cell lines reflecting the allelic diversity in field populations from which they were originally derived. Interestingly, at six loci, Teylovac™ alleles were present at a higher frequency in DNA preparations from vaccinated compared to unvaccinated Turkish cattle. The fact that immunising alleles were more frequently detected in this group initially suggested a direct effect of vaccination. However, the differences in frequency in the vaccinated and unvaccinated group were not statistically significant. Moreover, independent analysis of isolates from Sariköy village supported this finding. The results suggested that immunising with the Teylovac™ cell line had no detectable impact on genotypic diversity in the field. The lack of immunising alleles at particular loci in vaccinated animals suggested that either the genotype did not parasitise the red cell population or the allele from the immunising cell line was present only as a minor component, since this dataset only reflected the predominant allele. The absence of evidence of the immunising genotype contributing alleles to the field population was anticipated, since the immunising cell line was attenuated and so unlikely to differentiate to the merozoite. Consequently it was predicted to be absent from the piroplasm / erythrocyte fraction of the blood sample which was used for genotyping. Therefore, a more suitable isolate collection (i.e. schizont infected cell lines) may be required to definitively investigate this issue and determine if the cell line genotype can be detected instead in the leucocyte fraction. In addition, Turkish cattle designated as unvaccinated in this study may be more accurately described as cattle with no history of vaccination. It is therefore possible that a proportion of these animals were previously vaccinated with no record being taken. Therefore the results must be interpreted with an element of caution with future studies specifically designed to assess the impact of cell line vaccination on field populations. It is important to determine whether immunising stocks can contribute to erythrocyte infection and thus potentially recombine with field stocks to generate novel genotypes. This study supports the view that the Teylovac™ cell line does not proliferate in field populations and this is consistent with effective laboratory attenuation.

Futures studies should encompass vaccinated and unvaccinated cattle from the same site and ideally involve a distinct immunising genotype. The impact on population structure would need to be followed over the course of several years to determine the rate and level of recombination if present. Naturally, it would also be of interest to investigate whether immunising genotypes could be detected in ticks feeding from infected cattle.

CHAPTER FOUR

IDENTIFICATION OF MOLECULES UNDER POSITIVE SELECTION

4.1. Introduction

4.1.1. Identification of vaccine candidate genes

The advent of the fully annotated genome sequence of *T. annulata* (Pain *et al.* 2005) offers a novel opportunity to identify candidate genes encoding antigens for inclusion in a sub-unit vaccine to combat tropical theileriosis. The application of bioinformatics to interpret the vast amount of coding information in such an extensive dataset provides the means for achieving this goal (Fields *et al.* 1999). Comprising 3,793 predicted coding sequences (CDS), the primary challenge is to screen the *T. annulata* genome and identify a manageable subset of potential antigens for future evaluation in the field. For other pathogen species for which a genome sequence is available, a variety of approaches have been undertaken to identify vaccine candidate antigens. These studies generally employ a form of functional / immunological assay, whereby a considerable number of genes or their products are screened to test whether they stimulate an effective immune response. Naturally, this methodology depends on the availability of a test with both high specificity and sensitivity for detecting an appropriate reaction. However, in the most part, such assays tend to be labour intensive and may involve the cloning and expression of a large number of genes. For this reason, in many instances bioinformatic screening has been used as a primary, low-grade filter to reduce the size of the dataset being screened. Then, after a conveniently small set of antigen candidates has been identified by the functional assay, a more extensive bioinformatic characterisation of the selected genes may be undertaken before they are passed forward for evaluation *in vivo*.

The joint approach of genome mining for surface-associated proteins and subsequently screening these genes *in vivo* in a model host species has been undertaken in prokaryotes with some success (Maione *et al.* 2005). Group B Streptococcus (GBS) represents a major cause of life-threatening infection to neonatal humans and in the majority of cases, this infection is acquired through mother to baby transmission during delivery. In order to develop a broadly protective vaccine against GBS, the genomes of eight strains of this highly variable pathogen were analysed. This identified a ‘core’ 80 % of the genome that

is shared by all strains and a variable portion of 765 genes not present in all strains. These universal and variable sub-genomes were then mined to identify surface-associated / secreted proteins. Although this identified a large set of candidate genes (589), more than half of them were subsequently expressed as soluble His-tagged or Glutathione S-transferase (GST)-tagged fusion proteins. These purified proteins were used to immunise female mice, which were subsequently mated and the offspring challenged with one of six strains of GBS. One 'core sub-genome' antigen and three 'variable sub-genome' antigens were found to significantly increase the offspring survival rate with each antigen eliciting a response against more than one, but not all challenge strains. When the immunising gene was not present in the challenge strain, the antigen was not protective. However, in several cases protection was not achieved even although the challenge strain carried the antigen-encoding gene. This stimulated the study to test whether variable gene expression and / or variable surface exposure could have accounted for a lack of recognition. It was found that levels of surface expression, measured by antibody binding to viable bacteria, correlated with the protective activity of the antigen. Although this method did not produce a universal vaccine, when all four antigens were used to immunise mice the combination was found to be highly protective against a panel of twelve challenge strains. It can be concluded, at least for this study, that genuine surface-membrane exposure is likely to be critical to whether a predicted antigen is effective. Surface-accessibility was shown to vary from strain to strain and the study highlights the necessity to predict molecules with both high expression and with conservation of the functional signal and membrane-binding motifs. A combination of bioinformatics, high throughput expression and immunological screening has also recently been applied to a major causative agent for bacterial septicaemia and meningitis, *Neisseria meningitides* (Pizza *et al.* 2000). 570 putative surface-associated proteins were identified using a combination of motif signature database searches and homology to known surface proteins. Similar to the Streptococcus study, a large number of proteins (350) were subsequently cloned and recombinant proteins expressed. Seven surface-exposed antigens conserved between *N. meningitides* subgroups were selected for further evaluation as vaccine candidates. Similar approaches have been implemented for *Streptococcus pneumoniae* (Wizemann *et al.* 2001), *Chlamydia pneumoniae* (Montigiani *et al.* 2002) and *Porphyromonas gingivalis* (Ross *et al.* 2001).

There are additional methods by which a genome may be screened to identify candidate antigens. In bacteria, anomalies in guanine and cytosine (G+C) content have been used to identify loci acquired by lateral transfer from other species or even genera (Allan and Wren

2003). In a separate study on *N. meningitides*, the genome was found to contain two large regions of foreign DNA, two of which encode proteins involved in pathogenicity (Tettelin *et al.* 2000) and which have been suggested as potential targets for vaccine development. In the case of *M. tuberculosis*, the availability of genome sequence of two strains of a pathogen has offered a unique opportunity to derive vaccine candidates through bioinformatics (Cole *et al.* 1998). Since secreted proteins of *M. tuberculosis* have been associated with a protective immune response, SignalP and other software was used to identify secreted proteins for subsequent analysis using a pattern-matching algorithm that predicts T-cell epitopes (De Groot and Rappuoli 2004). This resulted in a set of 97 putative T-cell epitopes being selected for *in vitro* screening, from which twelve have been selected for inclusion in a multi-epitope tuberculosis vaccine (De Groot and Rappuoli 2004).

The genome of a parasitic protozoan, *T. cruzi* has also been the subject of an integrated bioinformatic and immunological screen in order to identify vaccine candidate genes. As in other eukaryotic cells, surface-exposed proteins are often anchored to membranes by covalent linkage to a glycosyl-phosphatidylinositol (GPI) domain. Generally, this is an uncommon modification and therefore the presence of this motif can be used as a highly discriminatory parameter in the bioinformatic screening process. The efficacy of GPI-anchored proteins to elicit protective immunity in *T. cruzi* has enabled EST and GSS (Genome Sequence Survey) databases to be screened for novel vaccine candidate genes (Bhatia *et al.* 2004). Around 2,500 sequences were mined to identify 19 sequences that exhibit characteristics of secreted / GPI anchored proteins. Eight of these genes were found to be highly conserved surface antigens that are expressed on multiple life-cycle stages of the parasite. These genes were cloned into eukaryotic expression plasmids and used for DNA immunisation of mice. Subsequent ELISA-testing of immune sera confirmed each of the polyclonal sera raised contained antigen-specific antibodies, the majority of which exhibited trypanolytic and agglutination activity.

To date, most studies have primarily used a bioinformatic process to identify a subset of genes within the genome, which are fed forward into an immunological assay. However, the use of motif prediction data alone represents a considerable under use of the wealth of information contained within the genomic sequence. Fortunately, various imaginative forms of bioinformatic analyses have been developed which may facilitate a more subtle approach to the identification of a limited number of high-quality candidate genes.

4.1.2. Screening for positive selection *in silico*

Variation in pathogen antigens may be accounted for by either (1) multiple allelic forms distributed across the species or (2) differential expression of multiple loci within an individual. Virtually all currently identified single-copy polymorphic antigens or variant multi-locus antigens are surface associated; furthermore, most polymorphic antigens are to be found on the invasive stage of the pathogen, whereas the majority of variant antigens are on the surface of infected cells or on extracellular pathogens (Conway and Polley 2002).

The frequency-dependent selection hypothesis for antigens predicts that as immunity develops against common alleles, individuals encoding rare antigenic types are less likely to succumb to the host immune response. Since immunity to tropical theileriosis is generally prolonged, the selective advantage of the rare allelic type may soon be lost as it becomes abundant in the parasite population. This in turn would allow novel rare alleles to expand until they too are recognised by a large proportion of the host population enabling yet more new variants to disseminate. The eventual result of this process would be the evolution and maintenance of antigenic polymorphism within populations of *T. annulata*. It may be hypothesised that as *T. annulata* and *T. parva* are closely related, evidence of this phenomenon of positive selection may be detectable across the species. Since more than 85 % of the genes in *T. annulata* have direct orthologues in *T. parva* (Pain *et al.* 2005; Gardner *et al.* 2005), a comparison of the DNA encoding these pairs of sequences is possible and may be used to quantify positive selection between genes. This positive selection between the orthologous genes may be accounted for by two mechanisms: (i) 'directional selection', with each parasite species adapting to its own individual evolutionary niche and (ii) 'diversifying selection' as a result of selective pressure from the bovine immune system through the process of frequency-dependent selection. For the purpose of identifying those antigen genes under the influence of this informative diversifying selection, 'directional' selection may be considered analogous to noise. Unfortunately, it is likely to be impossible to distinguish between this 'signal' and 'noise' when simply comparing single genomes from two separate species. In order to differentiate between these influences over any given gene, it is necessary to compare a panel of allelic sequences representing isolates of *T. annulata* alone.

A potentially attractive method of detecting positive selection within a single genome sequence was recently suggested, using an index of codon 'volatility' (Plotkin *et al.* 2004). This method is based on the simple hypothesis that if a particular protein has undergone an

excessive number of amino acid substitutions, then the DNA sequence will contain an over-representation of 'volatile' codons. The volatility of each codon is defined as the proportion of its 'point-mutation neighbours' that encode different amino acids. In other words, codon volatility quantifies the chance of the most recent mutation at any of its three nucleotides resulting in a change of amino acid. However, this methodology has been discredited due to its unjustifiable assumption that codon usage is shaped by selection for low codon volatility (Sharp 2005). The observation of unusual codon usage in candidate genes was instead attributed to mechanisms for generating the underlying repeat structure present within these genes. Zhang (Zhang 2005) suggests that other factors unrelated to positive selection such as codon bias relating to expression levels or to third position GC content may influence codon volatility. For these reasons, this method is believed to have limited value for detecting positive selection and was not used in this study.

Conway (Conway and Polley 2002) reviewed several methods, which have been successfully used to determine which areas of the genome are under selection. One method of demonstrating if a region of the genome is under positive selection is to compare the rate of non-synonymous substitutions (d_N) with the rate of synonymous substitutions (d_S). An area under positive selection would favour non-synonymous substitutions, resulting in a change of amino acid in the protein sequence. The utility of this method on a large scale to identify genes on which positive selection may operate was demonstrated in an early study (Endo *et al.* 1996). The d_Nd_S ratio of a large number of homologous sequences lodged in Genebank was calculated, identifying 17 groups of genes across a range of species as being under the influence of positive selection. Notably, nine of these groups were found to encode surface antigens of parasites and viruses. These include the merozoite surface antigen of *P. falciparum* (*MSA-2*), the major surface protein of *Anaplasma marginale* (*msp1α*) and the outer membrane protein of *Chlamydia* (*omp*). The results indicate that positive selection, favouring diversity of amino acid usage, is a common feature of surface membrane-associated proteins and that this may be detected across an extensive set of DNA sequences. The fact that just 17 groups were identified within a set of 3,595 genes (less than 0.5 %) means that comparative d_Nd_S analysis has the potential to be highly discriminatory. This particular study determined selection over the entire length of the gene, however, had the results included groups where only a region of a gene displays d_N greater than d_S, a much larger set of candidates would have been suggested (estimated at around 5 %). In order to quantify the distribution of positive selected sites across one of the candidate genes identified using global d_Nd_S values, a 20-codon frame sliding windows method was used to analyse two alleles of *MSA-2* (Endo

et al. 1996). Two regions were identified where the number of non-synonymous changes was high and synonymous changes were absent. The first of these regions corresponds exactly to known antigenic epitopes, and suggests that positive selection is acting on such epitopes to evade the host immune response. The authors allude to the fact that genome composition and codon usage may influence $d_N d_S$ analysis. The GC content of the genome of *P. falciparum* is only 18 % at the third base of each codon (Musto *et al.* 1995), suggesting that some mutational pressure is able to direct nucleotide substitutions toward adenine (A) and thymine (T). Examination of the genetic code reveals that such AT substitutions are very likely to be non-synonymous. Thus, it may be difficult to detect synonymous change, as synonymous mutations towards or between G and C in the third base position may easily revert to A or T. The pattern of synonymous codon choice in *P. falciparum* was the subject of a further study (Musto *et al.* 1999). Two major trends were found to determine codon usage in *P. falciparum*: (i) extreme genomic composition, with A and T predominating at the third position and (ii) heterogeneity among genes which, in turn may be related to gene expression. Correspondence analysis (described in Section 4.2.3.) was used to identify two subsets of genes composed of either highly or lowly biased codons and these subsets were putatively considered to be highly or lowly expressed genes. Usage of a subset of 20 codons was found to be significantly higher among the putatively highly expressed genes and these codons were designated as putatively optimal. A more recent study has highlighted increased usage of GC-rich non-synonymous codons in highly expressed genes in this parasite (Chanda *et al.* 2005). Additionally, non-synonymous sites of highly expressed genes are more conserved than those of low expression genes and for synonymous sites, the reverse is true. That is to say, at a particular site in a highly expressed gene where an amino acid is of a type which may be encoded by a single codon, that codon and amino acid may be conserved between allelic sequences. In contrast, in a gene expressed at a low level such a site would be more prone to codon mutation, invariably resulting in amino acid polymorphism between alleles. However, at a particular site in a highly expressed gene where an amino acid is of a type that may be encoded by several synonymous codons, there is greater amino acid polymorphism between allelic sequences at this site than observed in a comparable gene expressed at a low level.

In light of these studies, it must be appreciated that variation in codon usage both within and between *Theileria* species is likely to exist and may be related to levels of gene expression and other unknown factors. Therefore, in order for a genome wide $d_N d_S$ comparison between *T. annulata* and *T. parva* to be meaningful, codon usage across the

dataset must be taken into account as $d_N d_S$ may be heavily biased by non-standard codon usage.

4.1.3. An integrated bioinformatic approach

In *T. parva*, a protective immune response is known to involve recognition of peptides presented by MHC class I molecules on the surface of schizont-infected leucocytes by $CD8^+$ cytotoxic T-cells (Morrison *et al.* 1995). This cell type is also believed to play a role in immunity against *T. annulata* infected cells (Ahmed *et al.* 1989). Theoretically, the presence of T-cell epitopes on a protein may indicate that it is antigenic and one might predict that the identification of such motifs could be used to screen a large number of genes *in silico* for antigenicity. MHC class I epitopes have previously been bioinformatically predicted on conserved and polymorphic regions of the macroschizont antigen, TaSP (Schnittger *et al.* 2002). Of 22 putative epitopes, 14 were identified in areas on the protein that are conserved among allelic sequences representing different *T. annulata* genotypes. This study demonstrates that it is both worthwhile and interesting to predict epitopes in individual candidate molecules; however, there is limited software currently available for determining bovine T-cell epitopes. Although it would be possible to use software specifically designed for human MHC alleles to attempt to identify bovine epitopes, the value of such predictions is unclear. At the moment, therefore, the use of epitope prediction software in a screening process to identify *T. annulata* antigens must be regarded as speculative at best. In addition to the T-cell response, it is believed that macrophages may play an important role in developing long term immunity (Preston *et al.* 1999). Unfortunately, the nature of the parasite-specific stimulus for macrophage activation is unclear. Therefore, with current gaps in immunological knowledge and bioinformatic limitations, it is not possible to perform a highly targeted search to specifically identify the antigenic determinants involved in natural protective immunity. However, an *in silico* genomic screen which is based on other motif predictions may well be capable of identifying candidate antigens, which could be involved in natural protection.

A focussed bioinformatic approach may be more suitable for identifying sporozoite and merozoite antigen genes in *T. annulata*, since these are invasive extra-cellular bloodstream stages of the parasite and it may be expected that surface mounted proteins would be directly exposed to the bovine immune system. Several features could be used as selective criteria for identifying such proteins - (1) confirmation that the antigen is expressed by the extra-cellular stages, (2) secretion from the parasite and (3) tethering to the parasite membrane. As described in Section 1.3.1., the macroschizont is a complex intra-cellular

stage of the parasite. This is reflected in the current expressed sequence tags (EST) database, where a high number of genes are known to be expressed by this stage (1,407), contrasting with the lower number identified for both the merozoite (855) and piroplasm (838) stages (Pain *et al.* 2005). Unfortunately sporozoite EST data is currently unavailable. Merozoite proteins identified with this EST set, therefore only account for 23 % of the genome, representing a smaller more attractive set for data mining than the macroschizont genes, which account for 37 %. Although the EST dataset may be incomplete, genes encoding proteins that are in abundance on the surface of the merozoite would probably be highly expressed as mRNA and are therefore likely to be included in this EST set. Although non-classical pathways have been shown to allow secretion of proteins in *P. falciparum* (Nacer *et al.* 2001), previously identified surface antigen genes in *T. annulata* do encode a signal peptide, i.e. *TaMS1*, *SPAG* and *TaSP* (Katzer *et al.* 1994; d'Oliveira *et al.* 1996; Schnittger *et al.* 2002). TaSP has three transmembrane domains, which may be involved in tethering it to the surface of the schizont (Schnittger *et al.* 2002). Most notably, however, both SPAG and TaMS are each predicted to possess a glycosyl-phosphatidylinositol (GPI) anchor domain (GeneDB). Generally, such proteins lack a transmembrane domain, have no cytoplasmic tail and are located exclusively on the extracellular side of the plasma membrane. In protozoa, GPI anchors represent the predominant mechanism for integrating cell-surface proteins into the lipid bilayer (Turner 1994). Therefore, for both extra-cellular, invasive stage genes, a GPI-anchor was shown to be evident, while for the intra-cellular stage it was predicted that multiple transmembrane domains are used. This is strong supporting evidence that the combined presence of a signal sequence and GPI-anchor may be a signature of sporozoite or merozoite surface antigen status. Moreover, the rationale of targeting GPI-anchored proteins was successfully used in the study of *T. cruzi* discussed earlier (Bhatia *et al.* 2004).

From the results of the studies discussed in Section 4.1.2., it is clear that genomic analysis has been successfully used to predict vaccine candidates in other organisms and in the case of *M. tuberculosis*, comparative analysis of two genomes has aided vaccine design. The availability of the *T. parva* genome offers a unique opportunity to investigate whether interspecies diversity supports the prediction of a subset of genes as candidate antigens. For example, if GPI-anchored surface proteins are indeed antigens, they may be under the influence of positive selection by the bovine immune system and may be driven to diversify; comparative genomics analysis may provide evidence of this positive selection in the form of elevated interspecies $d_N d_S$ values for such genes where orthologues in each species can be identified. Consequently, an analysis of the distribution of $d_N d_S$ values

between different classes of gene is warranted to determine if this hypothesis is valid and to potentially discover other classes of antigen. However, as previously discussed, bias in codon usage, which is related to gene expression and other factors may influence $d_N d_S$ calculations. Therefore a codon usage analysis of the genome of *T. annulata* is needed to determine if such bias exists. In other words, comparative $d_N d_S$ analysis is required to support the motif identification approach and in turn a genome-wide codon usage analysis is required to validate the inter-species $d_N d_S$ approach.

4.1.4. Aims of this chapter

This chapter aims to answer the following specific questions -

- Can a manageable set of candidate antigen genes be identified using the bioinformatic methodology described in Section 4.1.3.?
- Do *T. annulata* / *T. parva* interspecies $d_N d_S$ values correlate with secretion and surface exposure in support of putative antigen identification?
- Does differential codon usage influence *T. annulata* / *T. parva* $d_N d_S$ distribution across the secretome and between life-cycle stages?

4.2. Materials and methods

4.2.1. *T. annulata* genomic resources

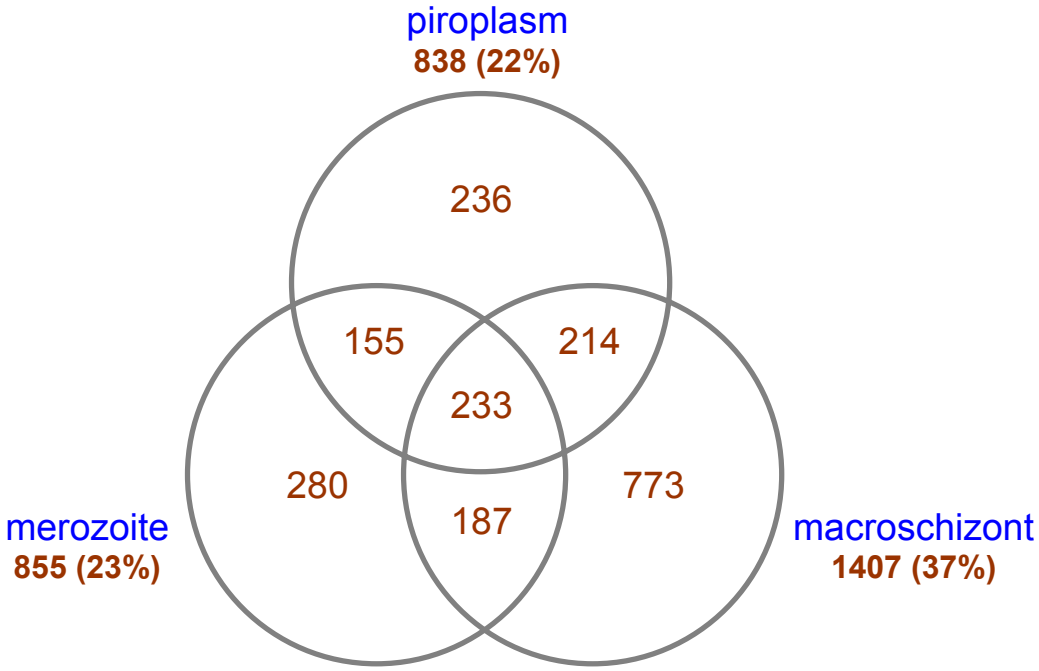
A summary of information held in the *T. annulata* genome database, publicly available at http://www.genedb.org/, was abstracted and supplied in June 2004 by Arnab Pain at the Sanger Institute. Updated EST information was incorporated into this dataset in January 2005 and it was subsequently used as the data source for much of the screening analysis. EST data was generated as described in the supplementary data accompanying publication of the *T. annulata* genome (Pain *et al.* 2005) and is summarised in Figure 4.1. Comparative genomics analysis was performed using the published sequence of *T. parva* (Gardner *et al.* 2005) with orthologous genes identified across the genomes using a reciprocal BLAST (Basic Local Alignment Search Tool), searching method. $d_N d_S$ values for each pair of genes were calculated using the CODEML feature of PAML (Yang 1997) using the basic codon model, M0. The variety of motif prediction software used to annotate the genome of *T. annulata* is described below.

Figure 4.1. Expressed sequence tag matches across three life-cycle stages of *T. annulata*

Expressed sequence tag (EST) data were generated for three life-cycle stages of the parasite, i.e. the macroschizont, the merozoite and the piroplasm. The merozoite dataset represented a differentiating macroschizont culture, and it is likely that a subset of these matches corresponded to genes expressed in the macroschizont. Matches were identified for a total of 2,078 genes across the 3,793 coding sequences (CDS) in the genome.

(adapted from Pain *et al.* 2005)

Figure 4.1. Expressed sequence tag matches across three life-cycle stages of *T. annulata*



4.2.2. Bioinformatic prediction of sequence motifs

4.2.2.1. Signal peptides

The SignalP2.0 HMM algorithm was utilised to identify proteins which enter the secretary pathway by virtue of encoding of a signal peptide (Nielsen *et al.* 1997). The software is widely acknowledged as accurately predicting the presence and location of signal peptide cleavage sites in amino acid sequences from eukaryotes. An improved version of the software based on a combination of several artificial neural networks and Hidden Markov Models, SignalP3.0 (Bendtsen *et al.* 2004), was used in the analysis of sub-telomeric variable secreted protein (SVSP) genes. An online implementation of this process can be found at http://www.cbs.dtu.dk/services/SignalP/.

4.2.2.2. Transmembrane protein topology

Transmembrane protein topology was also determined using an algorithm based on a hidden Markov model (Krogh *et al.* 2001). The software, TMHMM, is able to accurately discriminate between membrane and soluble proteins and is able to identify 97 – 98 % of transmembrane helices. A review of the variety of tools available to predict the topology of transmembrane proteins concluded that TMHMM is currently the best performing program (Moller *et al.* 2001). An online version of the software is hosted at http://www.cbs.dtu.dk/services/TMHMM/.

4.2.2.3. Glycosyl-phosphatidylinositol (GPI) anchoring

GPI-anchored proteins were identified using proprietary software, DGPI v. 2.04, which is documented online at http://129.194.185.165/dgpi/DGPI_demo_en.html. The software operates by first screening for the presence of a signal peptide at the N-terminus of a query protein using an adaptation of a previously published algorithm (von Heijne 1986). If this is identified then a novel filtration method is used to search for hydrophobic regions consisting of 13 amino acid residues at the C-terminus, which incorporates a cleavage site.

4.2.2.4. Nuclear localisation signal

A nuclear localisation signal (NLS) is a peptide motif that directs transport of proteins into the nucleus. Proteins containing these motifs were identified using PredictNLS software, which compares a query protein sequence with a set of known NLS. The process is also capable of identifying DNA binding motifs (Cokol *et al.* 2000). The online version, currently hosted at http://cubic.bioc.columbia.edu/predictNLS/, was used in this study.

4.2.2.5. Protein families

The Tribe-MCL protein cluster algorithm (Enright *et al.* 2002) was used to group proteins from *T. annulata* into putative families. The various problems associated with other protein sequence clustering algorithms such as the presence of multi-domain proteins and fragmented proteins are not encountered with this novel method, which utilises a Markov CLuster (MCL) algorithm to assign proteins into families based on pre-computed sequence similarity. This method has been successfully used to detect and categorise protein families within the draft human genome. The list of the top 30 *T. annulata* family clusters can be found in Table S3 of the online supplementary data which accompanied publication of the genome (Pain *et al.* 2005).

4.2.3. Codon usage analysis software

4.2.3.1. Indices of codon usage and gene composition

Codon usage was analysed using the software package CodonW, written by John Peden at the University of Nottingham. This software is distributed through a public license and can be downloaded from the internet at http://codonw.sourceforge.net/. The software calculates several standard indices of codon usage and gene composition and can be used to identify putatively optimal codons. These indices of usage are:

Relative Synonymous Codon Usage (RSCU) is the ratio of the observed frequency of a codon relative to that expected if codon usage is uniform, i.e. values tending towards one indicate an absence of bias. This measurement is independent of amino acid composition and is therefore useful for analysing codon usage in genes with a high level of internal repetitive structure, such as the SVSP family members.

The effective number of codons (ENc) is a simple measurement of codon bias representing the number of equally used codons required to generate the observed codon usage bias. For example, if a gene was extremely biased, and used a single codon for each amino acid, the value would be 20 and in a completely unbiased gene it would theoretically be 61.

GC_{3s} represents the frequency of guanine or cytosine nucleotides at the third position in synonymous codons.

GC skew (GC) measures the skew in the frequency of guanine and cytosine nucleotides and is calculated as the ratio of the difference compared to the sum of their respective frequencies.

Amino acid length (L_{aa}) is simply the number of amino acids than encode the hypothetical protein product.

Hydrophobicity score (Gravy) is an index of the average hydrophobicity of the encoded protein and is calculated as the arithmetic mean of the sum of the hydrophobic indices of each amino acid residue.

Aromaticity score (Aromo) is the proportion of amino acid residues that are aromatic across the entire translated gene product.

4.2.3.2. Correspondence analysis

CodonW was also designed to implement correspondence analysis (COA), a popular method of multivariate analysis. Correspondence analysis is a data ordination technique, which can determine the major trends in the variation of the data and may be used to distribute genes along continuous axes in accordance with identified trends. Such multivariate statistical techniques are necessary to summarise and explain the complex variation that may be encountered when analysing codon usage. The advantage of correspondence analysis, as opposed to cluster analysis, is that it does not make the assumption that the data is partitioned into discrete clusters and may therefore accurately represent continuous variation. CodonW calculates the first 40 axes describing the variation and indicates the quantity of variation explained by each axis (relative inertia). Results are plotted graphically using SigmaPlot (version 8.02).

4.3. Results

4.3.1. Comparative genomic $d_N d_S$ analysis

Genuine orthologues of 3,254 *T. parva* genes have been identified in the genome of *T. annulata*. This dataset was ranked in order of ascending $d_N d_S$ value and divided into seven class intervals of approximately equivalent numbers of genes and the proportion of each class that possessed a signal peptide was calculated. The results of this can be seen in Figure 4.2. Around 6 % of the lowest class ($d_N d_S$ 0.00 - 0.05) possessed such a motif rising to approximately 30 % of the highest class, which contained genes with a $d_N d_S$ value greater than 0.30. A clear trend was evident between these extremes.

Figure 4.2. Proportion of genes with signal peptide across $d_N d_S$ class

For each of the 3,254 genes where orthologues were identified in the genomes of *T. annulata* and *T. parva*, inter-species $d_N d_S$ values were calculated. This dataset was ranked in order of ascending $d_N d_S$ value and divided into seven class intervals of approximately equivalent numbers of genes. Using the SignalP2.0 algorithm, the proportion of each class that possessed a signal peptide was calculated.

Figure 4.3. Mean $d_N d_S$ across differentially expressed secreted genes

Using EST data, a proportion of genes predicted to be secreted were classified as stage-specifically expressed in the macroschizont (n = 162), the merozoite (n = 90) and the piroplasm (n = 53). Genes expressed in all three stages were regarded as constitutively expressed (n = 10). Mean $d_N d_S$ values for each of these stages were calculated along with their respective standard errors. The 217 genes with no corresponding EST data were also included in the analysis.

Figure 4.2. Proportion of genes with signal peptide across $d_N d_S$ class

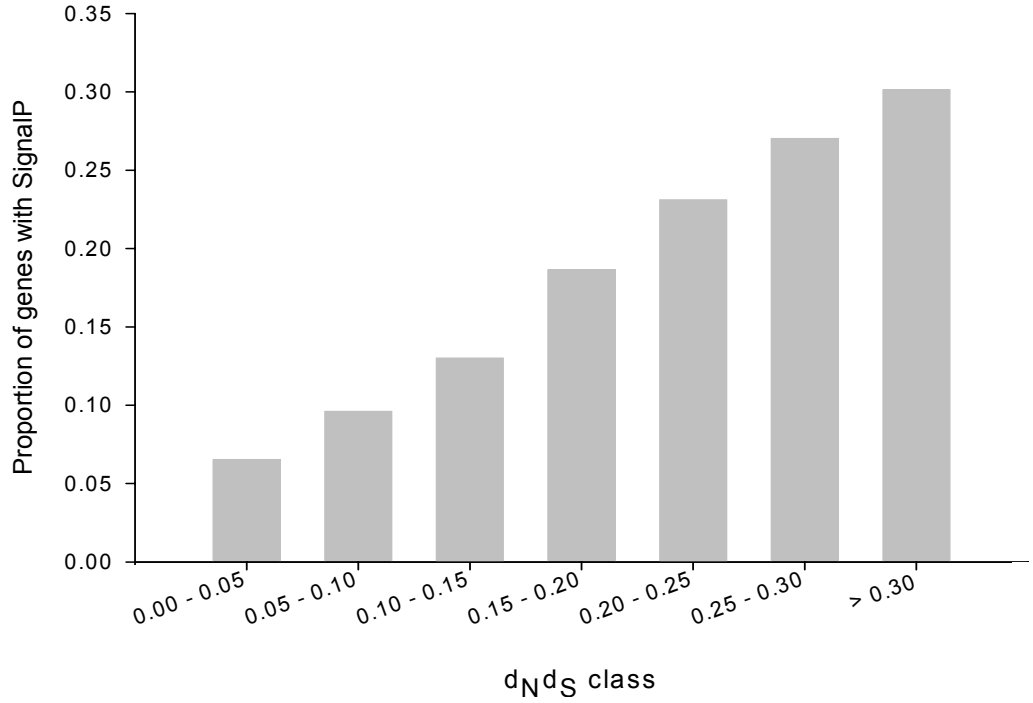
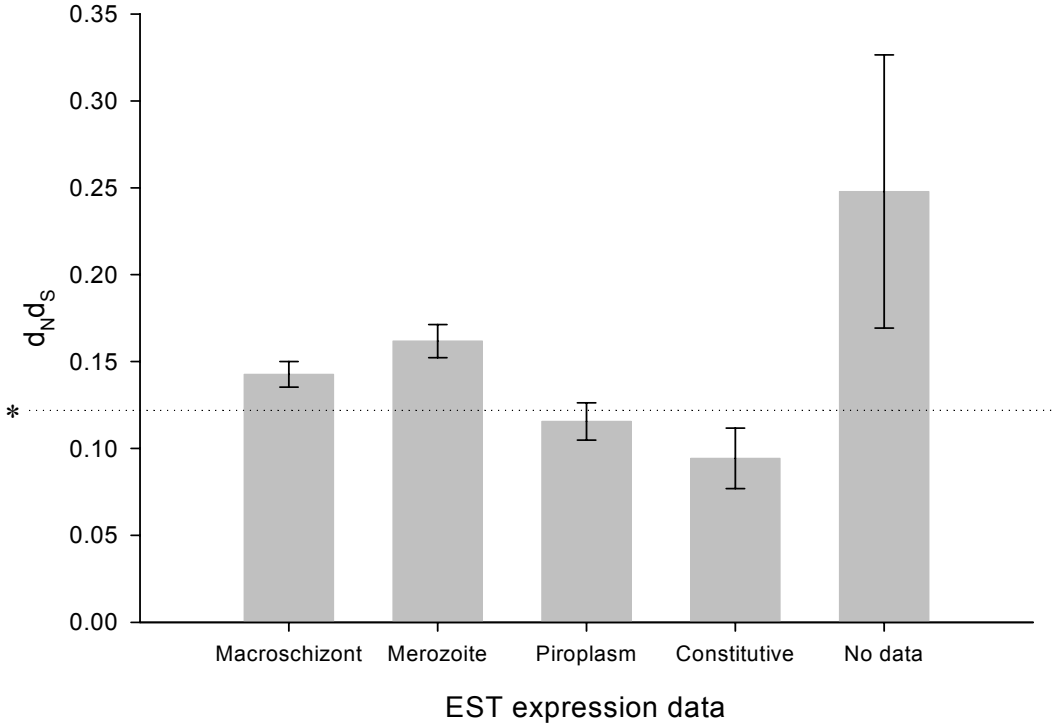


Figure 4.3. Mean $d_N d_S$ across differentially expressed secreted genes



* 0.1220, average $d_N d_S$ across all genes with orthologues

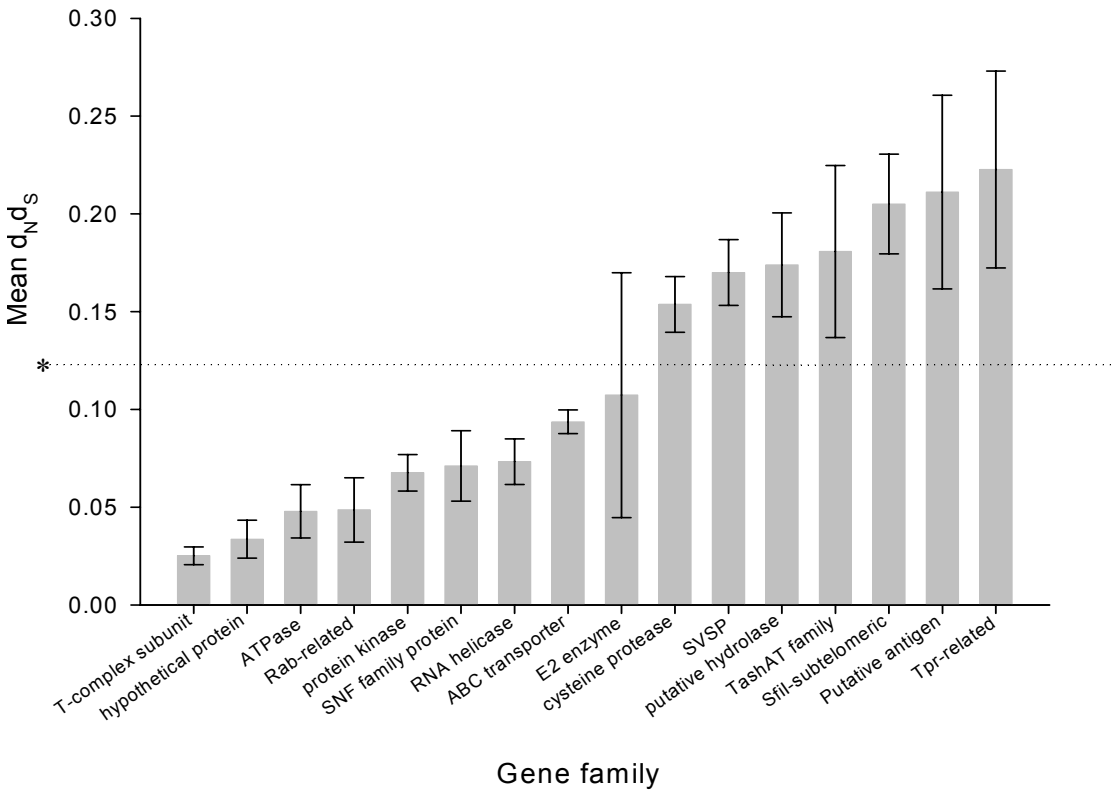
The 445 genes that possess a predicted signal peptide, based on analysis with SignalP2.0, are tentatively suggested as the *T. annulata* secretome. Since (a) the presence or absence of a signal peptide has a major correlation with the $d_N d_S$ value and (b) generally the genes of interest in the study are predicted to be secreted, the secretome was used as an unbiased dataset to attempt to quantify the distribution of $d_N d_S$ across differentially expressed genes. That is to say, by using a dataset that only includes genes encoding a signal peptide, this element of bias is removed from the analysis. EST data relating to macroschizont, merozoite and piroplasm life-cycle stages within the bovine host was used to classify genes within the secretome with respect to stage-specific expression. Genes expressed in all three stages were regarded as constitutively expressed. Mean $d_N d_S$ values for the genes in each expression category are depicted in Figure 4.3. along with their respective standard errors. The 217 genes with no corresponding EST data are also included. Interestingly, this group had the highest mean value of 0.248, however the large standard error of 0.0786 was consistent with its highly variable constitution. This is more than double the mean $d_N d_S$ value calculated across all orthologous genes, which is 0.1220. Piroplasm genes had a slightly lower than average value, while the lowest value (0.0943) was obtained for the limited number of genes that are constitutively expressed (n = 10). Both large groups representing macroschizont and merozoite genes had higher than average values and low standard errors (merozoite = 0.1620, n = 90; macroschizont = 0.1430, n = 162).

The distribution of $d_N d_S$ across 16 gene families was investigated to determine which particular families may be either conserved or which may be subject to directional or diversifying selection when compared with orthologues in *T. parva*. Tribe-MCL software bioinformatically identified 14 families in the genome of *T. annulata* that contained seven or more genes on the basis of shared motifs (Pain *et al.* 2005), irrespective of whether they encoded a signal peptide or not. This included a large family of secreted variable sub-telomeric proteins (SVSPs), which are expressed in the macroschizont stage. Although genome annotation initially suggested 48 members, BLAST searching using two typical SVSP sequences as queries suggested that the family is represented by 55 members. Typically SVSP genes comprise a single exon and the hypothetical gene product possesses a signal peptide at the N-terminus. The mean $d_N d_S$ value for each of these 14 families was calculated and is shown in Figure 4.4. Two additional groups were included in this analysis. The first comprised the previously described family of parasite-encoded host nuclear proteins, the TashAT family (Swan *et al.* 2001; Shiels *et al.* 2004). This 16-member family is relatively well conserved between *T. annulata* and *T. parva* with half of the genes having direct orthologues. The second group comprised antigens identified in

Figure 4.4. Mean $d_N d_S$ across putative gene families

Using Tribe-MCL software, 14 gene families were identified in the genome of *T. annulata* that contained seven or more members. The mean $d_N d_S$ value along with the standard error for each of these families was calculated and the families were ranked according to increasing mean $d_N d_S$. Two additional groups were included, (1) the TashAT gene family and (2) 'putative antigens', i.e. genes identified as antigens in previous experimental studies and those annotated as antigens in GeneDB. The analysis clearly separates the families into two types: a low $d_N d_S$ group consisting of 'housekeeping' gene families and a high $d_N d_S$ group, which includes several secreted gene families.

Figure 4.4. Mean $d_N d_S$ across putative gene families



* 0.1220, average $d_N d_S$ across all genes with orthologues

previous experimental studies and those genes annotated as antigens in GeneDB and this group was denoted as 'putative antigens'. It can be seen in Figure 4.4. that the gene families fall into two types – those with either below or above average $d_N d_S$. The nine families with below average $d_N d_S$ consist of enzymes such as ATPase, protein kinase, RNA helicase and E2 ubiquitin-conjugating enzyme and other types of molecule that one would not predict to be antigens. It also includes ABC transporters, which are integral membrane proteins coded in sub-telomeric regions of the genome. The variance of the $d_N d_S$ value of these gene families is relatively low, with the exception of the E2 ubiquitin-conjugating enzyme family that includes two genes with values of 0.2144 and 0.4434. The seven families with higher than average mean $d_N d_S$ values include cysteine proteases, a family of putative hydrolases and five families of genes that are *Theileria* specific. This agrees with a separate analysis, where genes annotated as 'conserved *Theileria* hypothetical proteins' were revealed to have a high mean $d_N d_S$ value (data not shown). The five gene families are – (1) SVSP proteins, (2) TashAT proteins, (3) SfiI sub-telomeric genes, (4) putative antigens and (5) Tpr-related proteins. SVSP genes had the lowest variance, with almost all members exhibiting a high $d_N d_S$ value. The two highest-ranking families were the putative antigens and *Tpr*-related genes with mean values of 0.211 and 0.223 respectively.

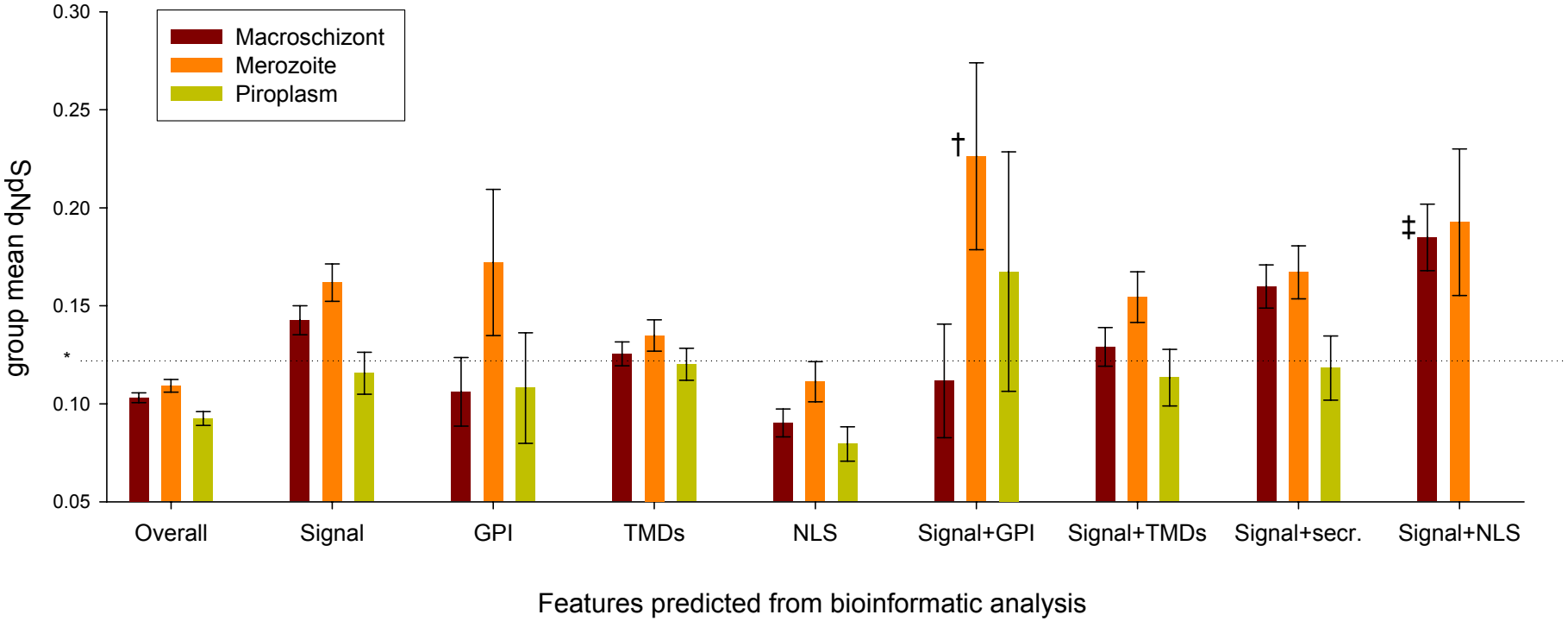
To test the hypothesis, described in Section 4.1.2., that genes with particular motif signatures may encode proteins under diversifying selection, a further $d_N d_S$ analysis was undertaken. EST and bioinformatic motif prediction data were integrated to determine mean $d_N d_S$ values across genes with a range of different predicted peptide motifs and expression profiles. The results of this analysis are presented in Figure 4.5. The presence or absence of a signal peptide (Signal), a glycosyl-phosphatidylinositol anchor (GPI), transmembrane domains (TMDs) and a nuclear localisation signal (NLS) were recorded to create both discrete and overlapping categories for mean $d_N d_S$ calculations. The column denoted 'signal+secr.' refers to a set of genes with a signal peptide but without either a GPI anchor or transmembrane domain.

The overall $d_N d_S$ values for each of the three life-cycle stages were below the mean value of 0.1220, suggesting that genes expressed in other stages of the life-cycle (e.g. sporozoites) have generally higher $d_N d_S$ values. This corresponded with the previously described analysis of the secretome, where genes with a predicted signal peptide and no EST data had high but variable values of $d_N d_S$. Proteins with a predicted GPI anchor had a noticeably high value in the merozoite stage contrasting with a low value in both the macroschizont and piroplasm. Presence of a TMD gave consistent values around the

Figure 4.5. Mean $d_N d_S$ across genes with variant predicted motif-signatures

EST and bioinformatic motif prediction data were integrated in order to determine mean $d_N d_S$ values across groups of genes with a range of different motif-signatures and expression profiles. The presence or absence of a signal peptide (Signal), a glycosyl-phosphatidylinositol anchor (GPI), transmembrane domains (TMDs) and a nuclear localisation signal (NLS) were recorded to create both discrete and overlapping categories for mean $d_N d_S$ calculations. 'Signal+secr.' refers to genes with a signal peptide but without either a GPI anchor or transmembrane domain. The Mann-Whitney test demonstrated that secreted merozoite proteins with a GPI anchor (†) and secreted macroschizont proteins with a NLS (‡) had a significantly different group mean $d_N d_S$. This test was used in preference to the t-test, because the values in each group were not normally distributed.

Figure 4.5. Mean $d_N d_S$ across genes with variant predicted motif-signatures



* 0.1220, average $d_N d_S$ across all genes with orthologues

† merozoite / signal / GPI proteins vs other merozoite proteins: *p = 0.016*, Mann-Whitney test

‡ macroschizont / signal / NLS proteins vs other macroschizont proteins: *p = 0.001*, Mann-Whitney test

average level for all three stages, while the presence of a nuclear localisation signal resulted in low $d_N d_S$ values, which were lower than the values in the overall dataset for macroschizont and piroplasm genes. Piroplasm expressed genes in all categories did not give $d_N d_S$ values significantly above the mean, suggesting that unlike *P. falciparum* (Kyes *et al.* 2001), no antigens are exposed to the immune system in the infected erythrocyte

An interesting finding was observed when the datasets corresponding to (a) genes with signal peptides and (b) genes with signal peptides and GPI anchors were compared. In the second group, the value for macroschizont genes falls below the average, while the value for merozoite and piroplasm proteins is higher. This observed high value for merozoite genes was tested for significance in comparison to merozoite genes that possess a signal peptide but no GPI anchor. As the values were not normally distributed, the Mann-Whitney test was performed, which was found to generate a *p* value of 0.016, confirming the significance of the difference between the groups. It should be appreciated that GPI anchors were identified by the DGPI v. 2.04 software only after the presence of a signal sequence had been established. However, this software used an alternative algorithm to the one used to identify signal peptides across the genome as a whole (SignalP2.0). This accounted for the difference in results between the 'GPI' genes and 'Signal+GPI' genes in Figure 4.5. Genes possessing a GPI motif but without a SignalP2.0 sequence may represent either an incorrect GPI prediction or the presence of a signal sequence not identified by SignalP2.0. The eight genes that possess a combination of signal peptide, GPI anchor and EST data for merozoite expression are listed in Table 4.1. When ranked in descending order of $d_N d_S$, four of the top five genes are annotated as antigens or putative antigens, the third ranking member being the major merozoite surface antigen of *T. annulata*, TaMS1. The other three comprise two orthologues of *T. parva* antigens and an orthologue of a *T. sergenti* merozoite / piroplasm antigen. The bottom three members of this group are generally more highly conserved with their *T. parva* orthologues than the top five members. Two of these are suggested as enzymes – an RNA helicase and a cysteine protease. The gene with the lowest $d_N d_S$ value of 0.0359 is hypothetical protein, which is highly conserved with *T. parva*. The multiple TMDs suggest this is an integral membrane protein consistent with the view that it may be a potassium ion channel, as implied by the Interpro and Pfam domain identification. The five top ranking members are further discussed in Section 6.4.

The presence or absence of a TMD in combination with a signal peptide did not noticeably alter the $d_N d_S$ value over any of the stages (Figure 4.5.). However, the absence of both

Table 4.1. *T. annulata* genes with signal peptide, GPI anchor and merozoite EST data

The eight genes that possess a combination of signal peptide, GPI anchor and EST data corresponding to merozoite expression were ranked in order of descending $d_N d_S$. These genes represented the merozoite 'Signal + GPI' group presented in Figure 4.5. '*T. annulata* ID' and '*T. parva* ID' correspond to the identification codes used in the GeneDB (www.genedb.org) and TIGR (www.tigr.org) genome databases. Four DUF domains were identified in TA08425, which is a domain of unknown function found in *Theileria* spp. Online databases, 'Interpro' (www.ebi.ac.uk/interpro/) and 'Pfam' (www.sanger.ac.uk/Software/Pfam/) had been used to annotate TA08325 as a putative ion channel. The top five ranking genes are considered to be putative antigen candidates and are highlighted.

Table 4.1. *T. annulata* genes with signal peptide, GPI anchor and merozoite EST data

T. annulata ID	*T. parva* ID	*T.a.* vs *T.p.*			EST data			GeneDB annotation	Transmembrane domains	Chromosome	Nucleotides	Introns	Amino acid residues	Mass (kDa)	Features
		Nucleotide identity (%)	Protein identity (%)	$d_N d_S$	macroschizont	merozoite	piroplasm								
TA16685	TP01_0987	73	62	0.4248		√		putative polymorphic antigen precursor-like protein	0	1	2937	no	978	110.7	Orthologue of *T. parva* 150 kDa microsphere antigen (p150)
TA08425	TP04_0437	68	58	0.2931	√	√		putative *Theileria parva* microneme-rhoptry antigen	1	4	2682	no	893	101.9	Orthologue of *T. parva* microneme-rhoptry antigen; 4 DUF domains (p104)
TA17050	TP01_1056	77	73	0.2751		√	√	merozoite-piroplasm surface antigen TaMS1	1	1	846	no	281	32.3	Well characterised merozoite & piroplasm antigen of *T. annulata*
TA20615	TP01_0487	77	71	0.256		√		hypothetical protein	1	1	1473	no	490	56.9	No homology with anything in database
TA13810	TP02_0551	85	83	0.1638		√	√	putative *T.sergenti* Chitose-type 23 kDa piroplasm surface-like protein	1	2	690	no	229	26.8	Homology with *T. sergenti* surface protein
TA16680	TP01_1165	78	77	0.1353	√	√		putative ATP-dependent RNA helicase	0	1	3088	yes	823	93.2	Homology with several RNA helicases
TA10955	TP04_0598	83	86	0.0904		√		putative papain-family cysteine protease	0	4	2087	yes	595	68.3	Homology with several cysteine proteases
TA08325	TP04_0417	81	91	0.0359		√		conserved hypothetical protein	12	4	4674	yes	1080	124.2	Putative ion channel (Interpro & Pfam domains)

TMDs and GPI anchors in secreted proteins correlated with an increase in $d_N d_S$ value in particular for the macroschizont stage. These proteins were predicted to be secreted by the parasite into its extracellular environment, 'Signal + secr.'. Another noteworthy finding in this analysis was that the presence of NLS in macroschizont proteins with a signal peptide significantly boosted $d_N d_S$. This group corresponded to some SVSP proteins and several members of the TashAT family, which were presented in Figure 4.4. When the Mann-Whitney test was performed comparing macroschizont / signal / NLS genes with macroschizont / signal / NLS negative genes a *p* value of 0.001 was obtained, confirming statistical significance. Interestingly, the graph suggested that merozoite proteins apparently share this property although the group variance is higher. However, it should be appreciated that merozoite EST data represented a culture that was undergoing asynchronous differentiation. Thus a gene, which was not turned off until late in the process of differentiation may not actually have been a 'merozoite' specific gene. This may account for high $d_N d_S$ of 'Signal+NLS' genes in the merozoite fraction. That is to say, they may not encode proteins secreted by the merozoite, but instead represent proteins secreted into the host cell by the macroschizont during the differentiation process.

4.3.2. Codon usage

The coding sequences (CDS) contained within the genome of *T. annulata* comprise 2,030,707 codons. Relative synonymous codon usage (RSCU) values for this dataset were calculated and are detailed in Table 4.2.(i). For each amino acid that may be encoded by more than one codon, it can be observed that certain codons are encountered more frequently than others. The codon with the highest RSCU value is AGA, which codes for arginine (Arg) with a value of 2.95. This figure contrasts with the lowest RSCU values for CGG (0.17) and CGC (0.29), two of the other five codons that encode this amino acid. For each of the other amino acids, which may be encoded by more than one codon, the value for the most frequent synonymous codon ranges between 1.18 for CAU (histidine) to 1.95 for UCA (serine).

To assess whether differential codon usage between *T. annulata* and *T. parva* may confound or invalidate interspecies $d_N d_S$ results, the RSCU of the 1,902,549 codons in the *T. parva* genome was also calculated (Table 4.2.(ii)). Similar to *T. annulata*, the amino acid with the most divergent RSCU was arginine with values ranging from 2.67 (AGA) to 0.24 (CGG). Also mirroring the situation in *T. annulata*, RSCU values ranging between 1.03 and 1.80 were seen for the preferred codons CAU and UCA. The general trend of RSCU towards particular codons was identical with the most frequent codon for each

Table 4.2. Relative synonymous codon usage

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

Table 4.2. Relative synonymous codon usage

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

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

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

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

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

(iii) *P. falciparum* genome (4,120,215 codons)

AA	codon	n	RSCU	AA	codon	n	RSCU	AA	codon	n	RSCU	AA	codon	n	RSCU
Phe	UUU	150152	1.67	***Ser***	UCU	60227	1.37	**Tyr**	UAU	209222	1.78	**Cys**	UGU	63344	1.74
	UUC	29393	0.33		UCC	21049	0.48		UAC	25724	0.22		UGC	9602	0.26
Leu	UUA	193620	3.73		UCA	68020	1.55	**TER**	UAA	4571	1.92	**TER**	UGA	1651	0.69
	UUG	43347	0.84		UCG	12463	0.28		UAG	905	0.38	**Trp**	UGG	20441	1.00
	CUU	35963	0.69	***Pro***	CCU	32280	1.57	**His**	CAU	85566	1.71	**Arg**	CGU	12488	0.69
	CUC	7463	0.14		CCC	8567	0.42		CAC	14363	0.29		CGC	1820	0.10
	CUA	24835	0.48		CCA	37207	1.81	**Gln**	CAA	98616	1.73		CGA	10033	0.55
	CUG	6216	0.12		CCG	3988	0.19		CAG	15235	0.27		CGG	1284	0.07
Ile	AUU	147749	1.17	***Thr***	ACU	43619	1.04	**Asn**	AAU	507563	1.72	**Ser**	AGU	85180	1.94
	AUC	25960	0.20		ACC	19777	0.47		AAC	82229	0.28		AGC	16145	0.37
	AUA	206552	1.63		ACA	89399	2.12	**Lys**	AAA	393405	1.63	**Arg**	AGA	65708	3.60
Met	AUG	90350	1.00		ACG	15586	0.37		AAG	88392	0.37		AGG	18038	0.99
Val	GUU	63333	1.60	***Ala***	GCU	33545	1.66	**Asp**	GAU	229865	1.73	**Gly**	GGU	48554	1.67
	GUC	10042	0.25		GCC	8634	0.43		GAC	35866	0.27		GGC	5586	0.19
	GUA	64828	1.64		GCA	34390	1.70	**Glu**	GAA	250684	1.71		GGA	50938	1.75
	GUG	19953	0.50		GCG	4489	0.22		GAG	42665	0.29		GGG	11536	0.40

(iv) *T. annulata* SVSP proteins (30,247 codons)

AA	codon	n	RSCU	AA	codon	n	RSCU	AA	codon	n	RSCU	AA	codon	n	RSCU
Phe	UUU	827	1.58	**Ser**	UCU	487	1.83	**Tyr**	UAU	1991	1.73	**Cys**	UGU	344	1.75
	UUC	220	0.42		UCC	76	0.29		UAC	309	0.27		UGC	50	0.25
Leu	UUA	737	2.57		UCA	532	2.00	**TER**	UAA	47	2.52	**TER**	UGA	6	0.32
	UUG	197	0.69		UCG	38	0.14		UAG	3	0.16	**Trp**	UGG	194	1.00
	CUU	315	1.10	**Pro**	CCU	1443	1.82	**His**	CAU	624	1.71	**Arg**	CGU	109	0.71
	CUC	51	0.18		CCC	101	0.13		CAC	106	0.29		CGC	19	0.12
	CUA	336	1.17		CCA	1569	1.98	**Gln**	CAA	1984	1.66		CGA	112	0.73
	CUG	82	0.29		CCG	61	0.08		CAG	409	0.34		CGG	17	0.11
Ile	AUU	1013	1.30	**Thr**	ACU	1054	2.10	**Asn**	AAU	1102	1.67	**Ser**	AGU	412	1.55
	AUC	143	0.18		ACC	101	0.20		AAC	215	0.33		AGC	48	0.18
	AUA	1178	1.51		ACA	803	1.60	**Lys**	AAA	1906	1.68	**Arg**	AGA	566	3.69
Met	AUG	407	1.00		ACG	53	0.11		AAG	368	0.32		AGG	97	0.63
Val	GUU	585	1.68	**Ala**	GCU	189	1.74	**Asp**	GAU	1505	1.76	**Gly**	GGU	488	1.35
	GUC	81	0.23		GCC	21	0.19		GAC	204	0.24		GGC	41	0.11
	GUA	645	1.85		GCA	209	1.92	**Glu**	GAA	2165	1.81		GGA	845	2.33
	GUG	84	0.24		GCG	16	0.15		GAG	232	0.19		GGG	75	0.21

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

amino acid corresponding exactly. The RSCU values in *T. annulata* for the more frequent codons were slightly higher and the values for the rarer codons were slightly lower than in *T. parva*. This is demonstrated graphically in Figure 4.6. where the RSCU of synonymous codons is correlated between *T. annulata* and *T. parva* and between *T. annulata* and *P. falciparum* (actual RSCU values of *P. falciparum* can be found in Table 4.2. (iii)). RSCU values of *T. annulata* are plotted on the x-axis while that of the other species is on the y-axis. Both axes are shown to intersect at 1, therefore codons in the upper or right hand sector are more frequent that those in the lower or left hand sector. Linear regression for the *T. annulata* / *T. parva* plot has a gradient that is close to, but less than, unity (m = 0.83). This underlines the general observation that preferred codons in *T. annulata* are also preferred in *T. parva*, but not to quite the same extent in *T. parva*. This stands in contrast with the relationship of codon usage of *T. annulata* in comparison with *P. falciparum*. Not withstanding a handful of outlying codons, the subset of preferred codons in *P. falciparum* is similar but their usage is even greater than in *T. annulata*, which is reflected in the linear regression line possessing a gradient of greater than one (m = 1.42). This is perhaps related to the greater AT richness of the coding sequences of *Plasmodium* compared with *Theileria*.

Correspondence analysis (COA) performed on the CDS of *T. annulata*, allows differentiation of genes based on codon usage. The first two axes generated in the analysis explain 18 % and 10 % respectively of the variation in codon usage. This can be seen in Figure 4.7., where the first four axes account for 40 % of the variation in codon usage. When the macroschizont/merozoite/piroplasm secretome is analysed, the axes are able to explain 22 % and 11 % of the variation. Subsets of the most highly and least biased genes may be identified as the 5 % of genes at either extreme of the principal axis. Where the RSCU for a particular codon is greater in the most biased compared to the least biased subset, a codon may be identified as putatively optimal. For the *T. annulata* genome, RSCU values for both subsets are presented in Table 4.3. Using this methodology, putatively optimal codons were identified across several datasets – the entire *T. annulata* genome, the entire *T. parva* genome, stage-specifically expressed *T. annulata* genes and secreted *T. annulata* genes with EST data. Summaries of the results are detailed in Table 4.4. Putatively optimal codons are broadly similar between both species, across the different bovine stages of *T. annulata* and within the subset of genes that comprise the *T. annulata* secretome. The putatively optimal codons identified by analysing the *T. annulata* genome as a whole are almost invariably identified across the different life-cycle stages and in the secretome, the only exception being `AUA`, `GUA`, `CCA` and `AGA` that

Figure 4.6. Correlation of relative synonymous codon usage of T*. annulata* & *T. parva* and *T. annulata* & *P. falciparum*

The relative synonymous codon usage (RSCU) of synonymous codons was correlated between *T. annulata* and *T. parva* and between *T. annulata* and *P. falciparum*, corresponding to the data contained in Table 4.2. RSCU values of *T. annulata* are plotted on the x-axis while those from *T. parva* (**red**) and *P. falciparum* (**green**) are on the y-axis. Linear regression lines for each of the comparisons are marked in **blue**, with a dotted **grey** line representing a perfect match, i.e. *T. annulata* vs *T. annulata*.

Figure 4.6. Correlation of relative synonymous codon usage of *T. annulata* & *T. parva* and *T. annulata* & *P. falciparum*

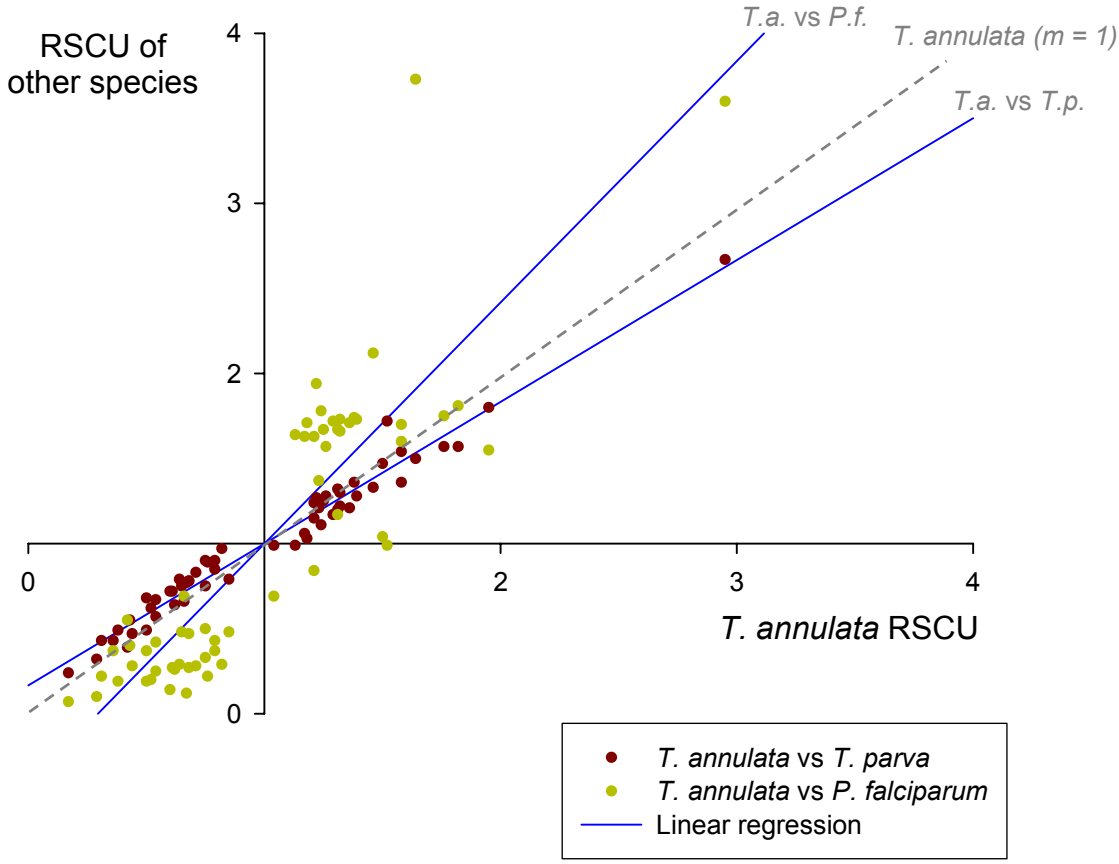


Figure 4.7. Relative inertia of axes from correspondence analysis of codon usage

Correspondence analysis (COA) was performed on the entire coding sequence of *T. annulata* in order to differentiate genes based on their codon usage. More than 70 % of the variation in codon usage is explained by the first twenty axes generated by the analysis, with the first four axes accounting for approximately 40 % of the variation.

Figure 4.7. Relative inertia of axes from correspondence analysis of codon usage

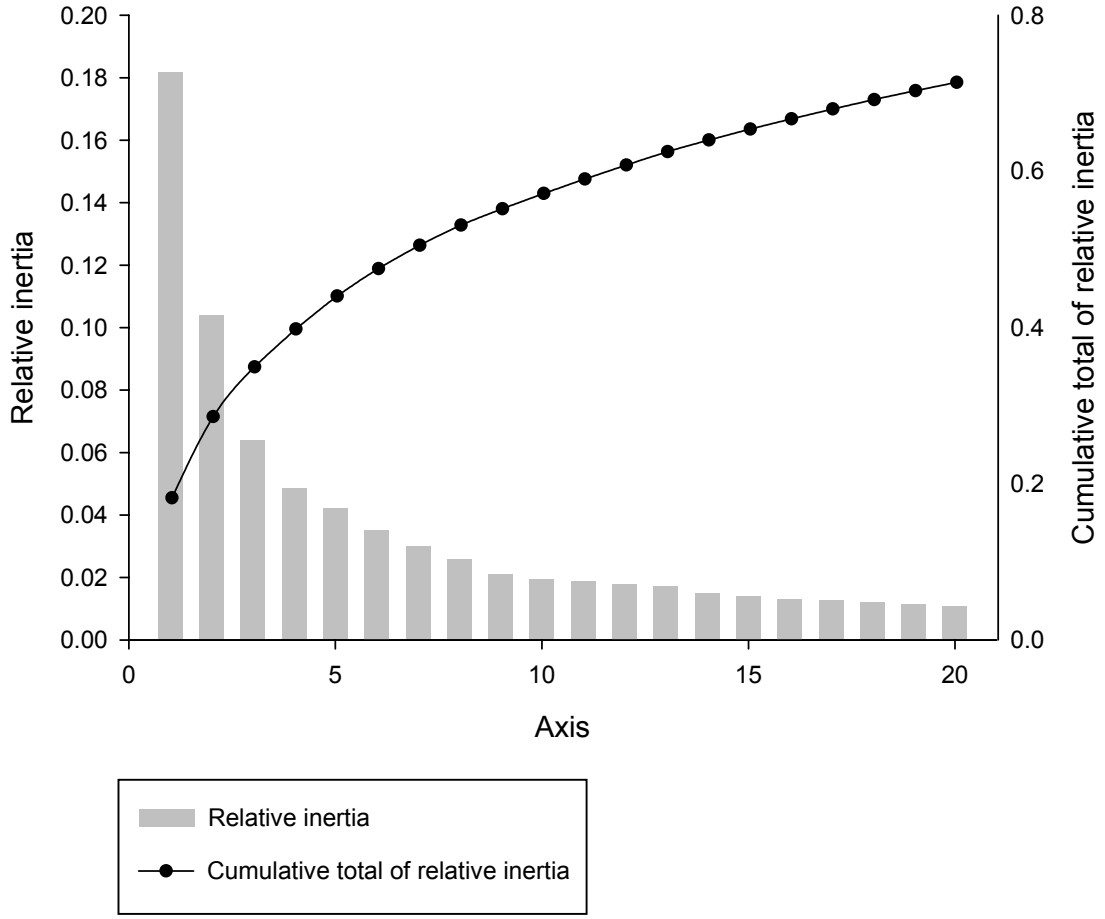


Table 4.3. Putative optimal codons of *T. annulata*

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

Table 4.3. Putative optimal codons of *T. annulata*

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

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

* $p < 0.01$

Table 4.4. Comparison of putative optimal codons of *T. annulata* and *T. parva*

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

Table 4.4. Comparison of putative optimal codons of *T. annulata* and *T. parva*

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

AA = encoded amino acid

√ $p < 0.01$, † $0.01 < p < 0.05$, * not statistically significant

are not identified in the piroplasm and CCA that is not identified in the merozoite. This probably reflects the smaller dataset for these stages. All optimal codons are identified in the macroschizont. Several additional optimal codons are suggested, however only six have statistical significance ($p < 0.05$). Interestingly, for five of these, where the optimal codon is not identified across the entire *T. annulata* genome, it is suggested as an optimal codon in *T. parva*.

To summarise, codon usage (as measured by RSCU) is almost identical in the genomes of *T. annulata* and *T. parva*. Moreover, the genomes share a subset of highly biased codons, which are largely invariant between life-cycle stages in *T. annulata*. Therefore, as there is no evidence of differential codon bias between the genomes of each species, meaningful comparisons (such as $d_N d_S$) can be made across the species and genes expressed in different life-cycle stages of *T. annulata* may be compared.

To determine what factors are involved in explaining the variation in codon usage, several indices of deviation in codon usage were calculated for each gene in the secretome with EST data. It was therefore possible to correlate these values with (i) 'ordination values' representing axes one through to four of the correspondence analysis and (ii) $d_N d_S$ values between *T. annulata* and *T. parva*. The resultant correlation co-efficients along with their statistical significance are detailed in Table 4.5. Three indices gave a strong negative correlation – the effective number of codons (EN_c), the frequency of G or C nucleotides at the third position in synonymous codons (GC_{3s}) and the GC skew (GC). That is, genes with high scores for each of these indices would tend to cluster towards the left hand sector and genes with low scores would tend to cluster to the right hand sector of COA plots. The Gravy score, which is based on the hydrophobicity of the translated gene product, was also negatively correlated with the value for axis one. Therefore the distribution of more hydrophobic proteins would also tend towards the left sector of the COA graph. The correlation for each of these four indices was statistically significant ($p < 0.01$). The GC skew and the Gravy score each correlated strongly with the second axis, although it was a positive correlation in the case of GC skew. Most importantly, $d_N d_S$ values correlated poorly with each of the axes from the COA with correlation co-efficients of 0.155 and 0.174 being recorded for axes two and four with *p* values less than 0.5 (axis two) and less than 0.01 (axis four). For axis one the value was 0.0851 and had no statistical significance. The relationship between axis one and EN_c, Gravy score and $d_N d_S$ is demonstrated graphically in Figure 4.8.(i, ii & v). The strongest correlation between $d_N d_S$ and any of the indices of codon usage deviation is with the Gravy score. The correlation is negative,

Table 4.5. Correlation of codon usage with potentially explanatory variables

Correspondence analysis of codon usage was performed on a dataset representing all *T. annulata* genes encoding a signal peptide and possessing EST expression data. In addition, for each gene in the dataset, several indices of deviation in codon usage were measured. Correlation co-efficients were calculated by comparing the first four axes generated by the correspondence analysis and gene $d_N d_S$ values with these various indices (Pearson's Product Moment correlation).

Axis *x*	represents the *x* th axis generated from correspondence analysis
EN_c	(the effective number of codons) represents the number of equally used codons required to generate the observed codon usage bias.
GC_{3s}	represents the frequency of guanine or cytosine nucleotides at the third position in synonymous codons.
GC	(GC skew) represents the skew in the frequency of guanine and cytosine nucleotides.
L_{aa}	represents the number of amino acids than encode the hypothetical gene product.
Gravy	(hydrophobicity score) is an index of the average hydrophobicity of the encoded protein.
Aromo	(aromaticity score) is the proportion of amino acid residues that are aromatic across the entire translated gene product.
$d_N d_S$	represents the non-synonymous to synonymous substitution rate (*T. annulata* vs *T. parva*).

Table 4.5. Correlation of codon usage with potentially explanatory variables

	EN_c	GC_{3s}	GC	L_{aa}	Gravy	Aromo	$d_N d_S$
Axis 1	-0.715 *	-0.937 *	-0.716 *	-0.038	-0.378 *	-0.089	0.085
Axis 2	0.049	0.026	0.561 *	0.108 †	-0.484 *	-0.255 *	0.155 †
Axis 3	0.185 *	0.252 *	0.292 *	-0.088 †	-0.166 †	-0.028	-0.053
Axis 4	0.154 †	0.148 †	-0.151 †	0.217 *	-0.709 *	-0.277 *	0.174 *
$d_N d_S$	0.012	-0.054	-0.018	0.174 *	-0.272 *	-0.063	

** $p < 0.01$*

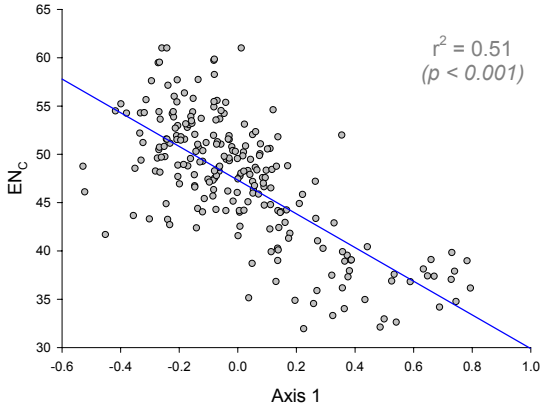
† *$0.01 < p < 0.05$*

Figure 4.8. Correspondence analysis results correlated with indices of codon usage

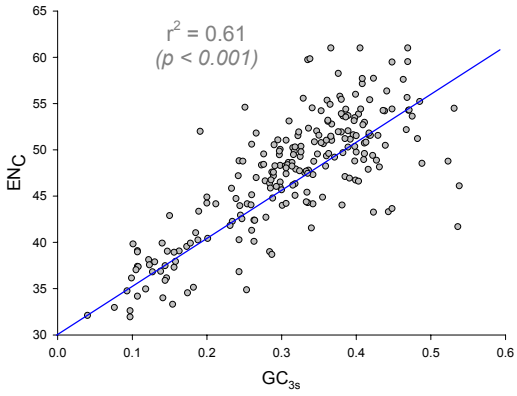
Indices of gene composition and codon usage were calculated for each gene in the secretome with EST data. The first axis generated by correspondence analysis of codon usage in this dataset (**Axis 1**) was plotted against the effective number of codons (**EN_c**), the hydrophobicity score of the translated gene product (**Gravy**) and the non-synonymous to synonymous substitution rate between *T. annulata* and *T. parva* (**$d_N d_S$**). Additionally, $d_N d_S$ was correlated with Gravy score. These four graphs correspond to correlation co-efficients described in Table 4.5. The relationship between EN_C and the frequency of guanine or cytosine nucleotides at the third position in synonymous codons (**GC_{3S}**) was also measured. Linear regression lines are marked in **blue**.

Figure 4.8. Correspondence analysis results correlated with indices of codon usage

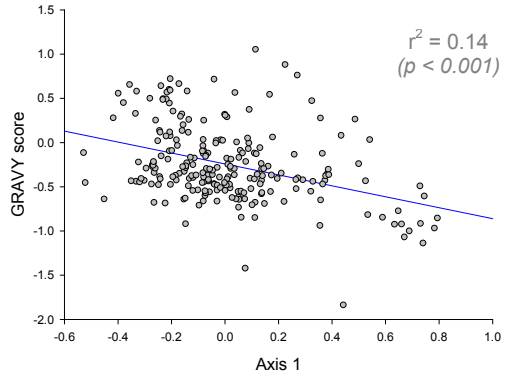
(i) Axis 1 vs EN_C



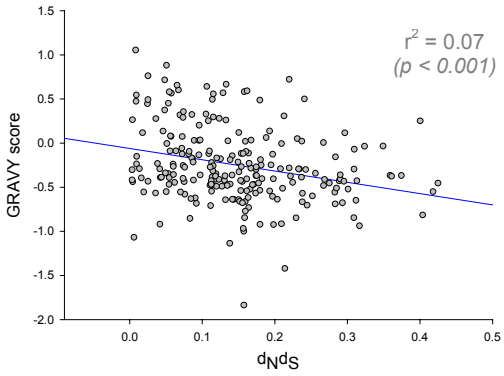
(ii) GC_{3s} vs EN_C



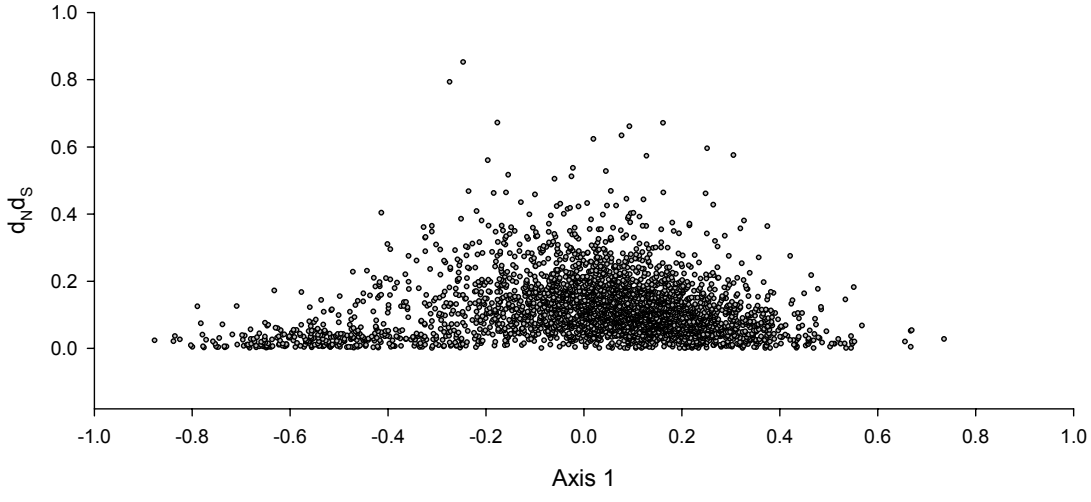
(iii) Axis 1 vs Gravy score



(iv) d_Nd_S vs Gravy score



(v) Axis 1 vs d_Nd_S



implying that gene products with higher hydrophobicity tend to have lower $d_N d_S$ values. This relatively poor but statistically significant correlation is shown in Figure 4.8.(iv). The relationship between GC_{3s} and EN_c is shown in Figure 4.8.(ii). Taken together, these results indicate that interspecies $d_N d_S$ values are not associated with indices of codon bias in the *T. annulata* genome.

To reveal whether stage of expression influences codon usage, the RSCU for non-synonymous codons was calculated for all genes with EST data, the results of which are shown in Figure 4.9. This clearly demonstrates that the preference of particular codons is almost identical across the three stages. Additionally, COA was undertaken using the dataset of genes displaying SignalP with expression in each single stage. The results of this analysis are presented in Figure 4.10. where the majority of genes are observed in a single cluster towards the centre of the graph. This cluster comprises most macroschizont and all piroplasm and merozoite genes. This indicates that in general, stage of expression does not explain the distribution of codon usage in genes encoding secreted proteins. A distinct group of genes can be seen centring on the lower left quadrant and comprising 13 macroschizont and one merozoite gene. When the annotation for these genes was examined, all the macroschizont genes were found to encode SVSP proteins and can be seen highlighted in the figure. Thus, SVSP proteins clearly deviate from the rest of the secretome in their codon usage. In order to investigate this deviation of the codon usage of SVSP proteins, RSCU was calculated and is detailed in Table 4.2.(iv). When compared with the RSCU of the *T. annulata* genome as a whole, generally there is a greater bias toward a set of preferred codons and this is shown graphically in Figure 4.11. Similar to the RSCU of *P. falciparum* compared with *T. annulata*, preferred *T. annulata* codons have an even greater value in the SVSP dataset compared to the genome as a whole. When linear regression analysis was performed the gradient of the line was 1.51. These results indicated that SVSP genes are encoded by a particularly biased subset of codons. This may simply represent a bias in amino acid composition or alternatively, if this subset represents *bona fide* optimal codons, it may be consistent with SVSP genes being relatively highly expressed.

Earlier analysis highlighted merozoite proteins with a signal peptide and a GPI anchor display high $d_N d_S$ values. In order to investigate how codon usage relates to this finding, COA was undertaken using the secreted proteins expressed in either the merozoite or the piroplasm stage. The results of this analysis are depicted in Figure 4.12., with the seven genes with a GPI anchor highlighted. These are distributed across the graph, and therefore,

Figure 4.9. Relative synonymous codon usage across stage-specifically expressed genes for non-synonymous codons

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

Figure 4.9. Relative synonymous codon usage across stage-specifically expressed genes for non-synonymous codons

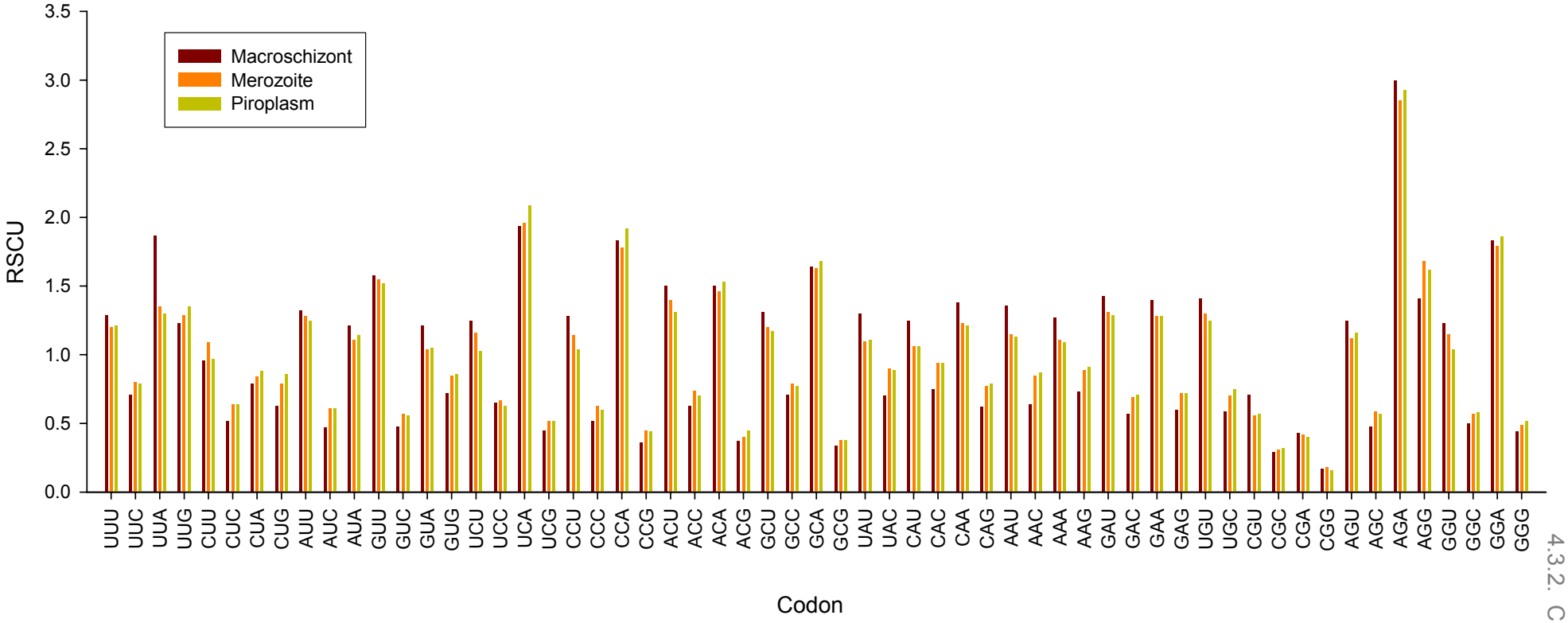


Figure 4.10. Correspondence analysis of codon usage of all genes with SignalP and stage-specific expression

A dataset comprising secreted, stage-specifically expressed genes was used to perform correspondence analysis of codon usage. The first two axes generated by this analysis were used to construct a graph. The proportion of the variation in the dataset explained by each axis is indicated in parenthesis. A cluster of SVSP genes with variant codon usage is highlighted, which also includes an outlying merozoite protein.

Figure 4.10. Correspondence analysis of codon usage of all genes with SignalP and stage-specific expression

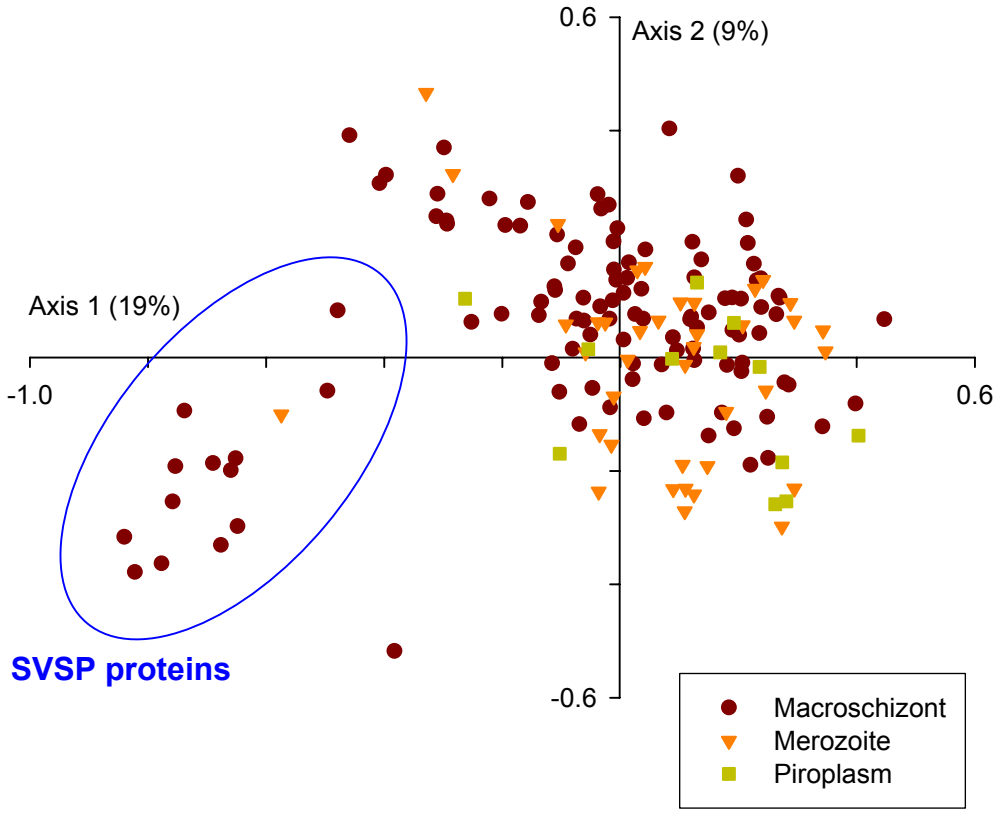


Figure 4.11. Correlation of relative synonymous codon usage of *T. annulata* & *T. parva* and *T. annulata* & SVSPs

The relative synonymous codon usage (RSCU) of synonymous codons was correlated between *T. annulata* and the subset of *T. annulata* genes representing the SVSP protein family. For comparison, a correlation between the genomes of *T. annulata* and *T. parva* is shown, corresponding to the data contained in Table 4.2. RSCU values of *T. annulata* (i.e. the entire genome) are plotted on the x-axis while that of *T. parva* (**red**) and *T. annulata* SVSP genes (**green**) are on the y-axis. Linear regression lines for each of the comparisons are marked in **blue**, with a dotted **grey** line representing a perfect match, i.e. *T. annulata* vs *T. annulata*.

Figure 4.11. Correlation of relative synonymous codon usage of *T. annulata* & *T. parva* and *T. annulata* & SVSPs

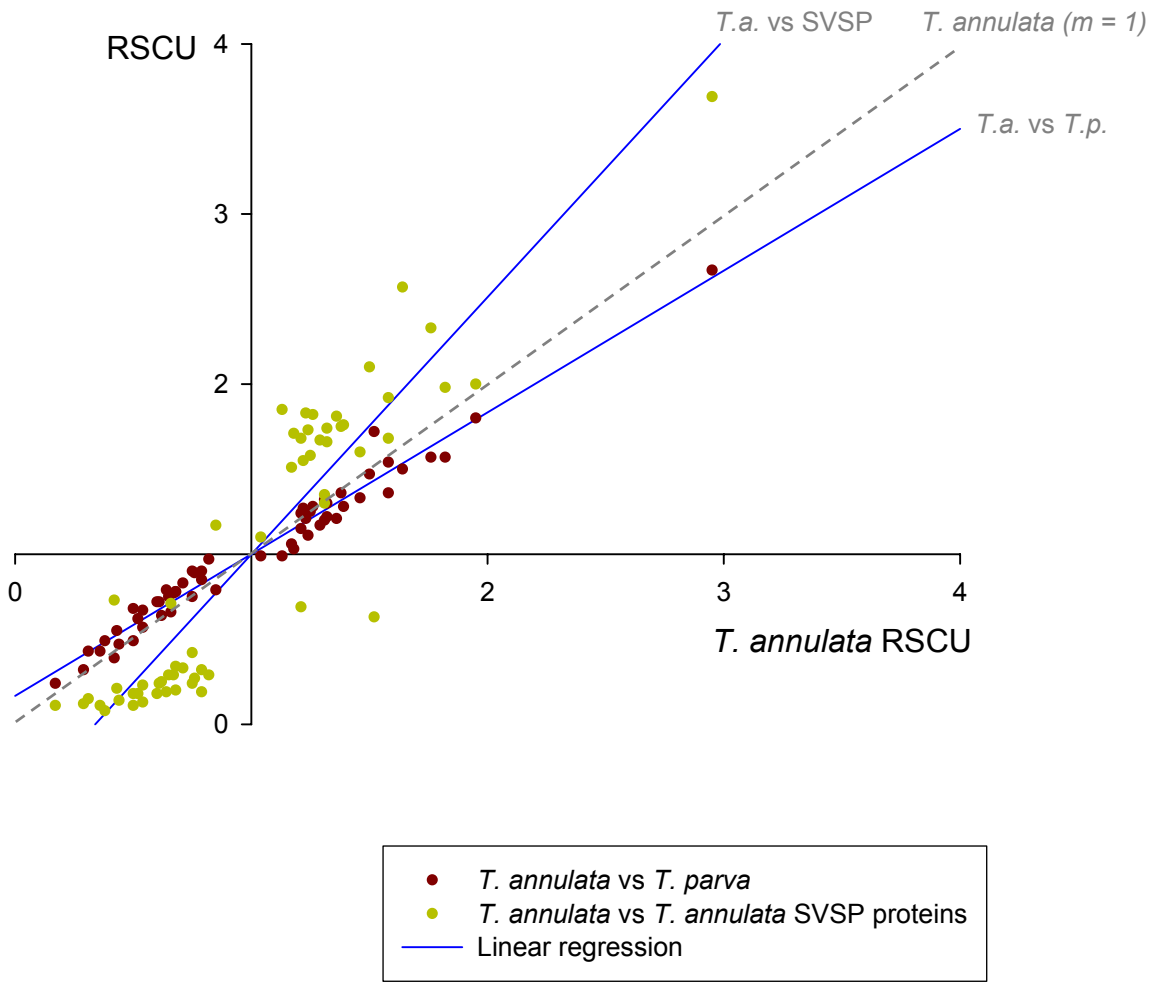
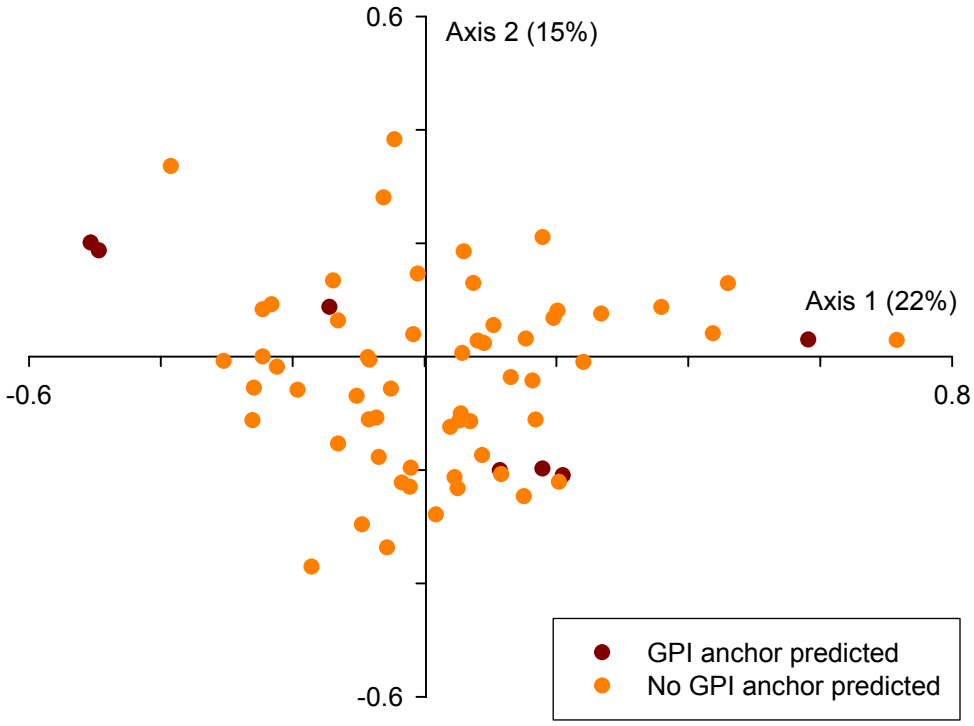


Figure 4.12. Correspondence analysis of codon usage of putatively secreted merozoite and piroplasm proteins

A dataset comprising genes with a signal peptide, which are stage-specifically expressed in the merozoite and piroplasm stages was used to perform correspondence analysis of codon usage. The first two axes generated by this analysis were used to construct a graph, which was labelled to denote the presence or absence of a GPI anchor motif.

Figure 4.12. Correspondence analysis of codon usage of putatively secreted merozoite and piroplasm proteins



most importantly, no general trend biasing codon usage can be identified across this subset of genes.

In order to discover an underlying reason for the bias in codon usage in SVSP genes, additional correspondence analysis was undertaken using three different measurements of gene variation across the entire secretome / EST dataset. COA of codon usage, amino acid composition and RSCU are depicted in Figure 4.13. COA of codon usage clearly shows the cluster of highly biased SVSP genes on the right hand side of the diagram (Figure 4.13.(i)). To determine if this pattern could be explained by unusual amino acid composition, the analysis was repeated using this measurement. A separate cluster of SVSP proteins is still evident in the lower right quadrant of Figure 4.13.(ii) suggesting that a biased amino acid composition may influence codon usage in the SVSP genes. When COA is performed using RSCU (Figure 4.13.(iii).), the SVSP genes continue to cluster together on the right. To summarise, a biased composition of amino acids combined with a biased RSCU account for the unusual codon usage observed in SVSP genes, whereas the merozoite proteins have a more 'standard' amino acid composition and codon usage.

To analyse codon usage and RSCU with respect to distribution of $d_N d_S$, the correspondence analysis graphs were reconstructed. The secretome was divided into classes of equivalent numbers of genes ranked according to $d_N d_S$ value. The codon usage and RSCU plots are shown in Figure 4.14. and are identical to the plots (i) and (iii) in Figure 4.13., but are colour-coded with respect to $d_N d_S$ class. The colour of the points representing SVSP genes in the cluster on the right side of the codon usage diagram reflects the relatively high $d_N d_S$ values that the group members possess. No general trend can be discerned from the rest of the data points. Similarly, when RSCU is analysed no obvious trend is observed with respect to $d_N d_S$ class. Therefore, it can be concluded that variance in $d_N d_S$ is not associated with either codon usage or RSCU (i.e. codon bias) across the secretome as a whole.

4.4. Discussion

4.4.1. General

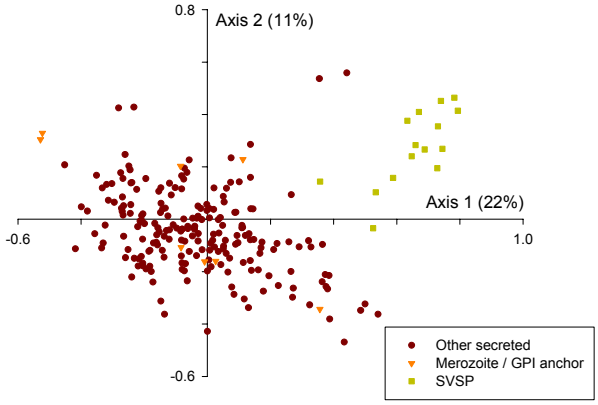
The work performed in this chapter was carried out in an attempt to mine genes from the *T. annulata* genomic database with bioinformatic signatures, which could identify them as putative antigens exposed to the immune system. Firstly, to investigate which gene families may be considered as putatively antigenic, a comparative genomic study was undertaken with the *T. parva* sequence, based on the premise that antigen genes exhibit elevated levels of $d_N d_S$. The results conclusively demonstrate that the presence of a signal

Figure 4.13. Correspondence analysis of secretome and proteins of interest

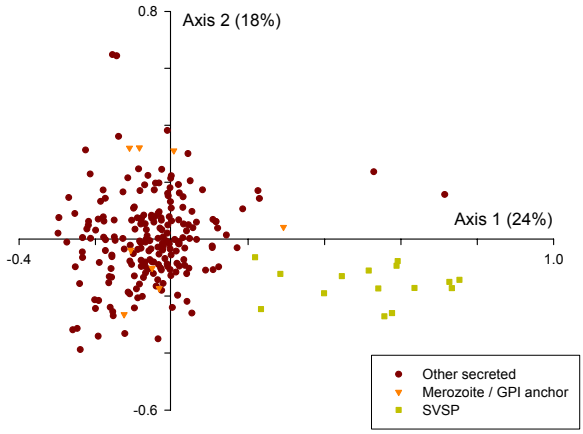
A dataset was compiled representing *T. annulata* genes that possess a signal peptide motif and are stage-specifically expressed. Correspondence analysis was performed on this dataset using measurements of **(i)** codon usage, **(ii)** amino acid composition and **(iii)** relative synonymous codon usage. The first two axes generated by each analysis were used to construct graphs, which were labelled to denote SVSP proteins **(green)**, merozoite proteins with a GPI anchor motif (**orange**) and other secreted proteins **(red)**.

Figure 4.13. Correspondence analysis of secretome and proteins of interest

(i) Codon usage (CU)



(ii) Amino acid composition



(iii) Relative synonymous codon usage (RSCU)

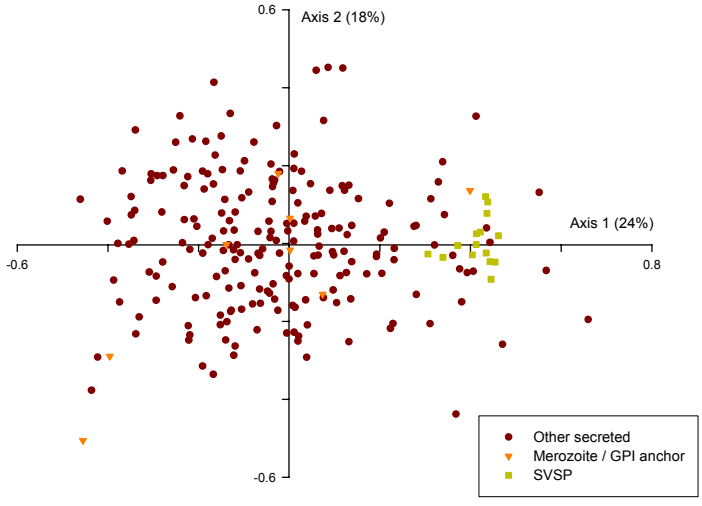
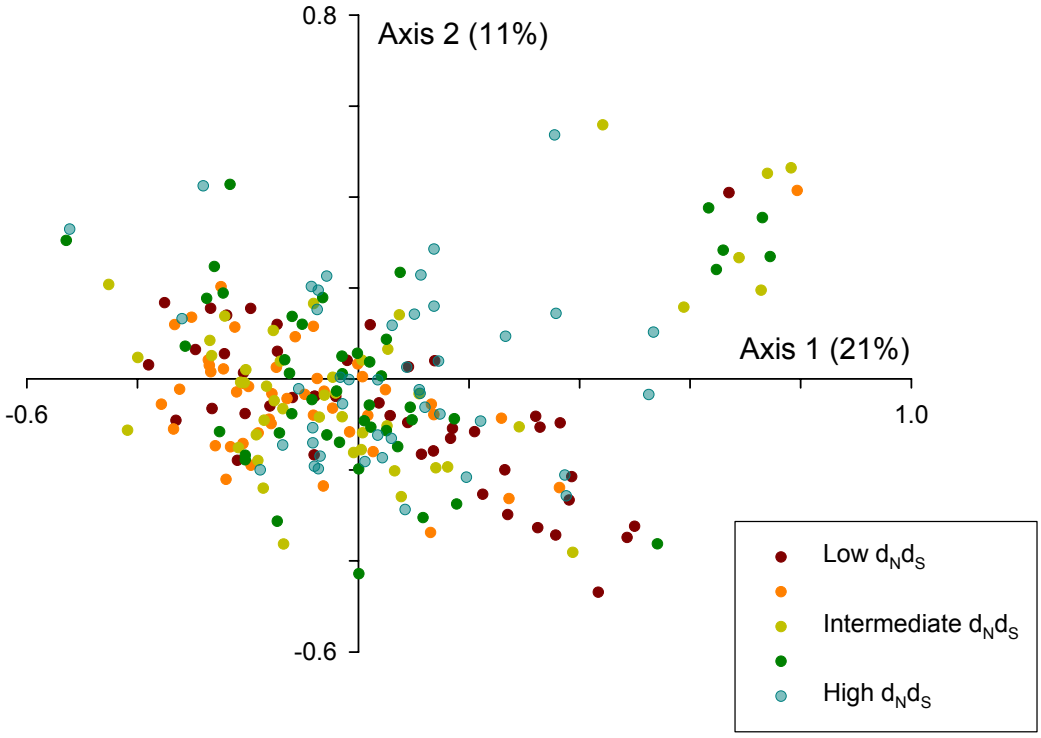


Figure 4.14. Correspondence analysis of secretome and $d_N d_S$ class

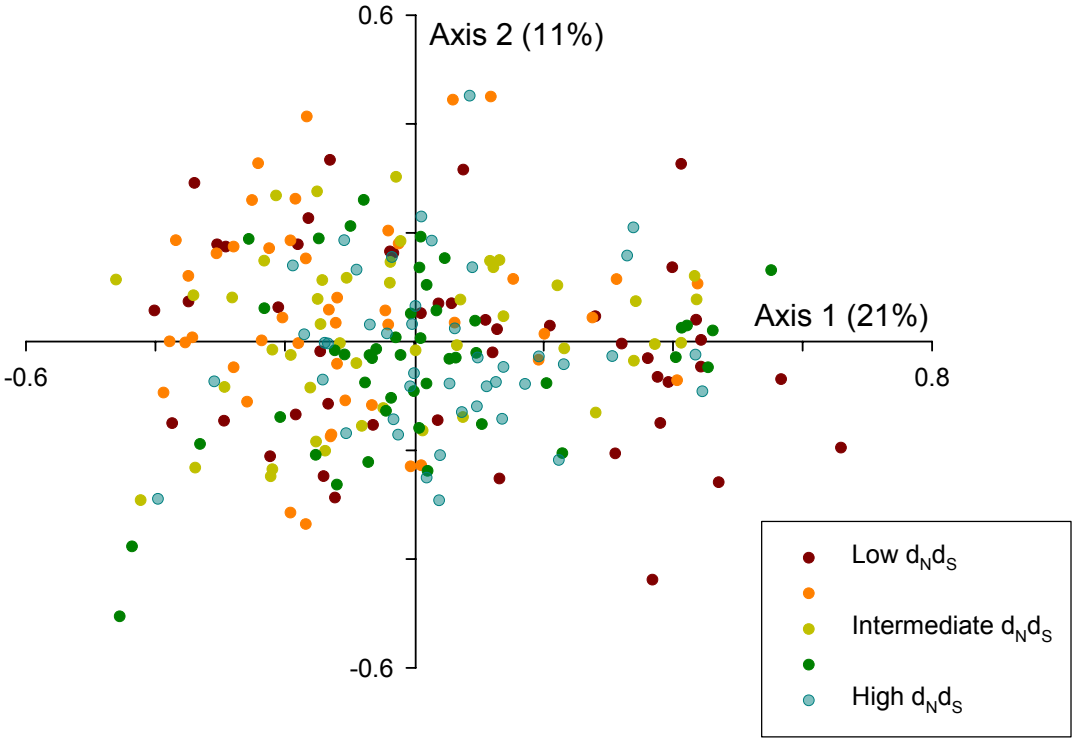
A dataset was compiled, representing *T. annulata* genes that possess a signal peptide motif and are stage-specifically expressed. Correspondence analysis was performed on this dataset using measurements of **(i)** codon usage and **(ii)** relative synonymous codon usage. The first two axes generated by each analysis were used to construct graphs, which were labelled to denote the $d_N d_S$ value of each gene. This was achieved by ranking the dataset in order of ascending $d_N d_S$ and separating it into five groups of equivalent size, which were differentially labelled.

Figure 4.14. Correspondence analysis of secretome and $d_N d_S$ class

(i) Codon usage (CU)



(ii) Relative synonymous codon usage (RSCU)



peptide is associated with genes of high $d_N d_S$, however it is remarkable that this trend is so apparent. 30 % of genes in the class with the highest value ($d_N d_S > 0.30$) possess a signal peptide compared to 6 % in the lowest class. The explanation for this may be rooted in the ancestry and function of the members of each group of proteins. The non-secreted portion of the proteome probably comprises a large number of housekeeping genes, performing critical functions within the parasite. Housekeeping genes would be predicted to be highly conserved, have an excess of non-synonymous substitutions and therefore would act to depress the mean $d_N d_S$ ratio. However, in *T. annulata*, the secretome may contain proteins and families of proteins involved in species or genus specific functions, e.g. proteins expressed on the surface of the parasite, some of which may be antigens. In evolutionary terms, such protein families may have evolved much more recently than highly conserved housekeeping genes common to the cells of many eukaryotic organisms. Therefore directional selection, where the parasite is progressively adapting towards its unique ecological niche, may be responsible for much of the observed signal sequence/$d_N d_S$ association. Consequently, to explore the relationship of protein families and $d_N d_S$ ratios, families of proteins based on Tribe-MCL clustering were identified and analysed. Prediction of protein families in large databases is one of the principal research objectives in functional genomics. Bioinformatic protein family classification can significantly contribute to the categorisation of functional diversity across related proteins and assist in the prediction of function based on domain architecture and sequence motif signature. The results for the protein families identified with this method (Figure 4.4.) clearly separates the families into two groups: a low $d_N d_S$ group consisting of housekeeping gene families (n = 9), a high $d_N d_S$ group which includes putative antigens genes and several secreted gene families. Two families from the latter group, SfiI and SVSP, are related protein families found at the telomeres. The highest-ranking group of proteins, designated 'Tpr-related' (also known as 'Tar' in *T. annulata*), is encoded by a multi-copy family of genes which was first identified in *T. parva* (Baylis *et al.* 1991; Bishop *et al.* 1997). The function of *Tar*/*Tpr* genes has not been elucidated, but their variable composition between genotypes has been used as a typing tool in *T. parva* (Stagg *et al.* 1994; Bishop *et al.* 1994a; Bishop *et al.* 1994b). These *Tar*/*Tpr* genes comprise the second largest family in the genomes of both species and unlike the SfiI and SVSP gene families, *Tpr* loci are not associated with the telomeres. The majority of *Tpr* genes form a single internal array on *T. parva* chromosome III (Gardner *et al.* 2005), however *Tar* genes are distributed throughout the four chromosomes in *T. annulata* causing minor breaks in synteny between the two species (Pain *et al.* 2005). Although the biological role of these loci are unknown, conservation in different *Theileria* species has indicated that they are functionally

important (Bishop *et al.* 1997). For most *Tar* genes, a direct orthologue cannot be identified in the genome *T. parva*. EST expression data indicates that of the 93 members currently identified in *T. annulata*, 51 are expressed by the macroschizont, merozoite and piroplasm together while 11 genes are expressed by the merozoite and piroplasm together. Additionally, four genes are piroplasm-specific; one is merozoite-specific while no EST data is available for the remainder. A small proportion of the family encodes a signal peptide and the vast majority of proteins display multiple transmembrane domains. This suggests that *Tar* genes encode a family of constitutively expressed integral membrane proteins. Although in many instances integral membrane proteins function as transporters, they may alternatively act as receptor molecules. If Tar proteins are located within the merozoite surface membrane, it is possible that certain domains are presented to the extra-cellular environment. In future studies it would be interesting to study the distribution of $d_N d_S$ along the length of these genes, to examine whether foci of high $d_N d_S$ corresponded to putatively exposed regions. Although high $d_N d_S$ is associated with immune selection, this alone does not necessarily indicate such genes are antigen candidates. The fact that few family members possess a signal peptide suggests that alternative mechanisms may be responsible for their high mean $d_N d_S$ value. Since *Tar* genes are highly variable, it is possible that high inter-species $d_N d_S$ could represent a degree of misalignment between orthologous genes in *T. annulata* and *T. parva*. Clearly, further research is required on this gene family to determine the sub-cellular location of their protein products and whether they may indeed encode antigens.

The $d_N d_S$ analysis provided a useful insight into the distribution of $d_N d_S$ among various gene families. However, to validate the results of this study, it was necessary to exclude the influence of codon bias from the analysis. This stimulated an investigation into codon usage across the genome of *T. annulata* and, to a lesser extent, the genome of *T. parva*.

4.4.2. Codon bias

It is clear that codon usage by *T. annulata* is not random, a common observation across the majority of eukaryotic species. It has been demonstrated in other organisms that the major factor explaining bias in codon usage between genes is the expression level of the encoded protein, with highly expressed genes using a limited subset of codons (Sharp and Matassi 1994). Discovering whether this is the case in *T. annulata* was not the purpose of this study and the role of codon bias in gene expression remains to be determined. This is likely to be addressed in the near future through (1) the availability of a more extensive EST dataset and (2) a planned genomic micro-array study designed to investigate gene

expression. Therefore, currently, the designation of putatively optimal codons remains speculative as it is based purely on excessive codon bias in a set of genes with as yet unquantified levels of expression.

The codon usage analysis presented in this study was primarily designed to detect whether biases in codon usage could explain $d_N d_S$ diversity. It is clear that $d_N d_S$ values correlate weakly or not at all with the amount of codon bias exhibited by individual genes. No relationship can be inferred between $d_N d_S$ and the principal axis generated during correspondence analysis and only a weak correlation is identified with the second and fourth axes (0.155 and 0.174). This can be accounted for by examining the relationship between the Gravy score and the second and fourth axes. Correlation co-efficients of -0.484 and -0.709 signify that genes encoding polypeptides with greater hydrophobicity will have a lower ordination value on that axis. Moreover, the strongest correlation of $d_N d_S$ with any of the indices is that with the Gravy score - a statistically significant negative correlation of -0.272. Hydrophobicity, effectively a measure of amino acid composition, is therefore the more important factor explaining distribution along these axes and $d_N d_S$ is consequently linked to this variable. This index of hydrophobicity has been previously used to explain the major trends in amino-acid usage in *E. coli* (Lobry and Gautier 1994). In that study, integral membrane proteins were discriminated from other proteins based on amino-acid composition using this parameter.

The genome of *T. annulata* has a mean GC content of 32.5 % and an even lower average figure in non-coding regions. Since base composition is believed to be a balance between mutational pressure towards or away from this pair of nucleotides, it is possible that the principal trend of codon usage in *T. annulata* may be explained by mutational bias such as variation in GC content. The fact that the indices of base composition, GC_{3s} and GC, correlate strongly with the principal trend supports this hypothesis. Additionally, the strong correlation of axis one with EN_c, an independent measure of codon bias, intuitively confirms that the correspondence analysis is providing a meaningful ordination of the genes with respect to codon usage.

Since (a) RSCU in *T. annulata* and *T. parva* is almost identical and (b) differential codon usage observed across *T. annulata* is not a major factor influencing $d_N d_S$ distribution among genes, it is reasonable to conclude that $d_N d_S$ is independent of the major trends in codon usage in this study. This allowed the comparative genomic $d_N d_S$ results to be regarded with some credence, stimulating further analysis of a number of groups of gene, which displayed relatively high $d_N d_S$ values. This included a class of gene putatively

encoding merozoite surface antigens and two previously described macroschizont gene families, the SVSPs and TashATs.

4.4.3. Merozoite antigens

From Figure 4.5., it can be seen that the merozoite proteins with a signal sequence generally seem to provide evidence that diversity has been positively selected. Furthermore, there is a synergistic relationship between the presence of a GPI anchor and a signal peptide in this stage, adding weight to the hypothesis that surface bound molecules are under strong selective pressure and are accumulating polymorphism as a means of evading the host's defenses. In contrast, presence of a GPI anchor or transmembrane domain correlates with conservation in the macroschizont expressed proteins, suggesting they are not under any pressure from the bovine immune system that is detectable using this methodology. The $d_N d_S$ results therefore suggest that parasite-encoded surface anchored proteins (i.e. on either the parasite or leucocyte surface) are not under the influence of positive selection in the macroschizont. This would imply that macroschizont antigens are devoid of the signature features (apart from a signal peptide) and that identifying a small subset of candidate genes from this dataset is not feasible.

Eight genes encoding predicted merozoite surface proteins were identified in the study. Results for merozoite $d_N d_S$ together with this limited number of candidate genes suggest members of this group offer a good opportunity for identifying antigenic proteins by bioinformatic mining of the *T. annulata* genome sequence. This conclusion for the merozoite dataset agrees with the findings of several *P. falciparum* studies (Hughes & Hughes, 1995), which demonstrated a high ratio of d_N to d_S in regions of sporozoite and merozoite surface protein encoding gene sequences. In support of these findings, a study of natural selection in *Plasmodium* (Escalante, 1998) showed that loci encoding proteins expressed on the surface of the sporozoite and the merozoite were more polymorphic than those expressed during the sexual stages or inside the parasite. This was in agreement with the general observation that stage-specific surface proteins exhibit high polymorphism when compared with internal antigens (McCutchan *et al.* 1988, Riley *et al.* 1994). Consistent with the Escalante study, the highest $d_N d_S$ figure for the *Theileria* dataset was for the extracellular merozoite stage, where the protein may be anchored to the surface and hence directly exposed to the host immune system.

It was encouraging that of the five top ranking genes, annotation of four suggested they encode antigenic proteins (Table 4.1.), with the other gene denoted as a hypothetical

protein. The two top ranking genes were orthologous to *T. parva* predicted antigens and were found to be between 2.7 and 2.9 kb in length. The *T. annulata* major merozoite / piroplasm antigen TaMS1 was identified in third position. This molecule has already been demonstrated to be highly immunogenic and has shown promise as a vaccine candidate (see Section 1.10.3.). A high level of diversity across alleles of this gene has been well documented over many parasite isolates (Gubbels *et al.* 2000b). Immune selection is currently considered to explain the intra-species polymorphism and the levels of $d_N d_S$ exhibited by this molecule. However, there is no clear evidence to date that genotypes displaying variant TaMS1 have a selective advantage during the bovine phase of the life-cycle. Interestingly, a preliminary study indicated that selection of TaMS1 variant parasites may occur following transmission through ticks, the hypothesis being that since the antigen is the major surface protein of the piroplasm stage that is taken up by a feeding tick, antibodies against TaMS1 in the blood of the bovine may impede further development in the tick gut (Gubbels *et al.* 2001). The orthologue of a merozoite / piroplasm surface antigen of *T. sergenti* (Sugimoto *et al.* 1991) was also identified (TA13810). As may be expected, this molecule is expressed in the merozoite and piroplasm stages of *T. annulata*. This has the highest nucleotide and protein identity of all five top candidates indicating a high level of conservation between the species. The fact that the gene is well conserved and that it also displays a relatively high $d_N d_S$ value is an observation that is rather counter-intuitive. However, it must be remembered than a high $d_N d_S$ value does not necessarily correlate with a high degree of sequence polymorphism. For example, the liver stage antigen-1 in *Plasmodium* (*lsa-1*) has limited amino acid polymorphism, however it does display relatively high $d_N d_S$ (Conway and Polley 2002); it is the quality of the sequence differences that is important, not the quantity. The encoded protein is 23 kDa in the Chitose strain of *T. sergenti*, but is a little larger in *T. annulata* at a predicted mass of 26.8 kDa, however it still much smaller than the top two candidates as it is encoded by only 690 nucleotides. A hypothetical protein (TA20615) was identified that does not have significant similarity to any other predicted products in the *T. annulata* genome. A search of peptide motif databases revealed no currently classified domains except the signal and GPI sequences. The nucleotide and amino acid identity with its *T. parva* orthologue is similar to that of TaMS1. The size of the predicted protein is calculated to be 56.9 kDa, intermediate in size between TaMS1 and the orthologues of *T. parva* microsphere antigen and the *T. parva* microneme-rhoptry antigen.

It is interesting to note from Table 4.1. that none of the top five genes are spliced, while the three conserved / low $d_N d_S$ genes all have introns. The significance of this finding is

difficult to assess, however, in higher eukaryotes it has been noted that intron-containing and intron-less versions of otherwise identical genes can exhibit dramatically different expression profiles (Nott *et al.* 2003). It was demonstrated that versions of genes with introns exhibited higher levels of gene expression than those encoded by a single exon. The presence of introns is also known to facilitate alternative splicing, whereby different regions of mRNA are removed before translation. The five top-ranking genes are likely to have an invariant structure since they are encoded by a single exon. The three low $d_N d_S$ genes have considerable homology to previously described non-*Theileria* proteins. Therefore, during the process of annotation, their structure may have been known *a priori*, allowing an accurate version of the gene model to be created in each case. However, it is unlikely that an inaccurate gene model accounts for the lack of introns in the high $d_N d_S$ genes. Since three of the five genes have been characterised in other *Theileria* species and since *TaMS1* has been studied in detail in *T. annulata*, the gene models are likely to be correct. The fact that the top five genes are probably *Theileria*-specific, leads to the supposition that they have evolved relatively recently and that this may be associated with their single exon structure. Another class of *Theileria*-specific genes with high group mean $d_N d_S$, whose members generally do not exhibit introns, is the TashAT family.

4.4.4. Parasite encoded host nuclear proteins

It is known that *T. annulata* secretes proteins from the macroschizont that locate to the host cell nucleus and it is, therefore, possible these are degraded in the host cell cytoplasm and exported to MHC molecules on the leukocyte surface (Swan *et al.* 1999; Swan *et al.* 2001; Swan *et al.* 2003; Shiels *et al.* 2004). To investigate the genes encoding these proteins, it was decided to include in the $d_N d_S$ analysis a category of predicted proteins that exhibit nuclear localisation signals. When found in combination with a signal peptide in stage-specifically expressed proteins, a large average $d_N d_S$ value indicated that secreted, non-membrane bound proteins may be the best candidates for evidence of antigen selection expressed by the macroschizont. This group includes the previously described TashAT family of proteins, known to be secreted by the macroschizont and locate to the host cell nucleus (Swan *et al.* 1999; Swan *et al.* 2001). Four of the nine genes with orthologues in the *T. parva* genome are predicted to contain a nuclear localization signal. Similarly, when considered as a family (Figure 4.4.), this group has a high mean $d_N d_S$ value. It would be of great interest to discover the underlying reason for this observation. Since *T. annulata* and *T. parva* have different biological features such as their host-cell tropism, differences between the orthologous gene families may due to directional selection and may be directly related to speciation. An alternative explanation is that the high $d_N d_S$ values may

be because the proteins are indeed exposed to the immune system and are subject to diversifying selection. Since TashAT family members are secreted by the macroschizont into the host cell where they migrate to the nucleus, it may be argued that these proteins are particularly at risk of degradation and presentation on the surface of the leucocyte via the MHC pathway. Furthermore, many of these TashAT genes contain PEST motifs, which are potential proteolytic cleavage sites. The motifs are defined as hydrophilic stretches of at least twelve amino acids with a high local concentration of the critical amino acids proline (P), glutamic acid (E), serine (S) and threonine (T), the presence of which considerably reduces the half-life of a protein. Interestingly, over the PEST domains, TashAT proteins show weak identity to a related family of genes, which also exhibit high $d_N d_S$, namely the SVSP family.

4.4.5. SVSP proteins

A large family of hypothetical proteins has been identified in sub-telomeric regions of the genome (Pain *et al.* 2005) using protein family prediction software (Tribe-MCL). 48 members were identified in *T. annulata* with 85 identified in *T. parva*, the largest identified family in that species. A cursory examination of these genes reveals many members share a number of striking features: (1) stage-specific macroschizont EST data, (2) a signal peptide sequence and (3) multiple PEST motifs. The further observation that none of these proteins contain a GPI-anchor motif suggests that they are stage-specifically expressed in the macroschizont and secreted into the host cell cytoplasm. Due to the presence of PEST motifs, it can be inferred that SVSP proteins may be targeted for rapid degradation in the host cell cytoplasm, similar to the TashAT family.

The high $d_N d_S$ values associated with this family of proteins might reflect regulatory functions that have diversified after speciation of *T. annulata* and *T. parva*. Alternatively, they might reflect exposure to the immune system, after rapid degradation to generate peptides, which are presented by major histocompatibility complex antigens on the infected cell surface. The presence of PEST motifs would support this view. The immune response to the infected cell will depend on the recognition of non-self polypeptides presented on the surface. If the predictions of the bioinformatic analysis are correct (i.e. secretion of the SVSP proteins into the host cytoplasm) coupled with the data showing high levels of macroschizont-specific expression of many SVSP family members, then the infected cell will present a wide range of non-self proteins on its surface. In principal, this would lead to a wide range of cellular responses with different specificities, assuming peptides of each SVSP family member has an equal probability of being presented on the

surface. In the case of a cytotoxic T-cell response, any lineage specific to a particular peptide would need to recognise a relatively non-abundant peptide on MHC Class I amongst a mosaic of many other related parasite peptides. This could, in principal, reduce the effectiveness of T-cell binding and recognition. Before taking such a hypothesis further, experimental evidence is required to establish whether SVSP peptides are presented on the surface of the infected cell and whether most family members are presented simultaneously.

This protein family was also independently highlighted in the course of the codon usage study. The subset of SVSP genes, which possess a signal peptide, macroschizont-exclusive stage-specific EST data and have a direct orthologue in *T. parva* were included in the codon usage analysis (Figure 4.10.). All 13 of these SVSP genes formed a cluster distinct from the bulk of the genes. It was initially suggested that this was due to the fact that these are related genes with repetitive glutamine and proline motifs and that therefore codon usage is biased towards a subset of codons encoding these residues. This was found only to partially account for this difference. When the secretome was analysed with respect to its RSCU, the SVSP genes still clustered at one extreme of the principal axis (Figure 4.13.(iii)). The underlying reason for this phenomenon may be that the SVSPs are composed of a highly biased set of codons (see Figure 4.11.) and that this bias may be associated with high levels of expression in the macroschizont. Supporting evidence of this theory may be found in the current EST data set; the majority of family members can be detected in the macroschizont, suggesting that individual expression levels must be appreciable. However, the differential codon usage of the SVSP genes from the rest of the secretome was shown to be partly due to unusual amino acid composition (Figure 4.13.(ii)) and this in turn may be related to the presence of low complexity PEST domains. The influence of these factors on the comparative value of interspecies $d_N d_S$ is difficult to assess. It is possible that codon bias within these genes contributes to the relatively high $d_N d_S$ identified in this gene family (Figure 4.4.). Conversely, it could be argued that similar composition (codon usage and / or amino acid composition) in the *T. parva* SVSP family largely neutralises any influence on $d_N d_S$. To further investigate whether the family is under positive selection it would be necessary to compare the allelic sequence of SVSP genes within *T. annulata* alone.

The multi-copy nature of the gene family and the fact that they may have evolved as contingency genes suggests that the polypeptides they encode are likely to be unsuitable as macroschizont vaccine candidates. Nevertheless, the bioinformatic approach has been able

to suggest the TashAT family genes as candidate macroschizont antigens, as previously discussed. Furthermore, a recent study in *T. parva* (Graham *et al.* 2006), conducted at the same time as this research, has highlighted the usefulness of applying bioinformatics to identify macroschizont vaccine candidate antigens from genomic sequence.

4.4.6. Antigen identification in *T. parva*

Preliminary sequence from the *T. parva* genome project has been mined in a combined bioinformatic and immunological approach to identifying vaccine targets for East Coast Fever (Graham *et al.* 2006). The goal of that study was to identify novel macroschizont antigens, since a protective bovine immune response has been shown to be mediated by cytotoxic T-lymphocytes (CTL) that lyse macroschizont-infected lymphocytes (Graham *et al.* 2006). The schizont lies free in the cytoplasm of the infected bovine cell, hence, proteins incorporating a signal peptide sequence would be predicted to pass into the MHC class I processing pathway for presentation on the surface of the infected cell. However, a study in *P. falciparum* demonstrated that a non-classical secretory pathway may operate in protozoa, which does not require the presence of a signal peptide (Nacer *et al.* 2001). Therefore, since the presence of signal peptide was a likely, although not a mandatory feature of secreted molecules, the study in *T. parva* comprised two approaches to immuno-screening - (1) with the subset of genes on chromosome I possessing a signal peptide and (2) with a random mixture of schizont cDNA clones. This methodology ultimately identified a panel of five candidate antigens, which were recognised by the CTL response. However, the first of these approaches involved cloning a set of 55 candidate genes and eventually identified only two out of the five genes as recognised by cytotoxic T-cells. Therefore, in this example, the bioinformatic criteria for targeting a set of genes identified a relatively large set of candidates and so lacked specificity and lacked sensitivity, as it failed to identify the majority of the genes identified by the random cDNA screening. Only the genes on a single chromosome were subjected to bioinformatic screening; screening the entire genome may have resulted in the identification of the three other genes. In part, this approach is analogous to the one taken in several prokaryotic species and in *T. cruzi*, discussed in Section 4.1.1., whereby a large set of genes are passed forward for functional assaying following a low-specificity *in silico* screen. In contrast, the approach presented in this thesis used a more discriminatory bioinformatic methodology. In addition to providing a broad overview of the families that may be under positive selection and which may function as antigens, various bioinformatic parameters were used to target a limited number of putative merozoite antigens. It must be appreciated that the presence of non-classical secretory mechanisms in *T. annulata* may have resulted in a

number of genuine secreted proteins being excluded from the analysis. However, in the case of merozoite candidate antigens, it would be a relatively minor task to revisit the dataset and identify the limited number of GPI-anchored proteins without a motif recognised by SignalP2.0. to augment the list.

The limited EST data currently available restricted the life-cycle stages that could be analysed to the macroschizont, the merozoite and the piroplasm. It is possible that with additional EST data, sporozoite candidate antigens may have been identified using similar criteria to that used for predicting merozoite antigens. It is likely that future micro-array experiments will characterise the expression profile of this critical extra-cellular stage. Furthermore, it is likely that the expression profile of the transient trophozoite stage is distinct from that of both the sporozoite and the macroschizont, as the parasite establishes itself in the host cell. Expression data relating to this stage may assist identifying novel candidate antigens expressed within the host leucocyte. However, the limited EST data currently available was used to predict a conveniently small number of candidate antigen genes. Moreover, additional bioinformatic techniques may be used to reduce this subset further, and provide a very limited panel of promising genes.

4.4.7. Further analysis

A further study is required to clarify the reasons for high $d_N d_S$ in the TashAT, SVSP and merozoite antigen candidate genes. It is necessary to quantify allelic diversity in *T. annulata* to assess whether the shaping force for high interspecies $d_N d_S$ is directional or diversifying selection for each group. This is particularly important in the case of the merozoite antigen genes, where the effect of frequency dependent positive selection needs to be determined before particular candidates can be selected for future *in vivo* studies. Consequently, a combination of methods is required to analyse species-specific sequence diversity for representative genes from each of these three groups. The results from such an approach are presented in Chapter Five of this thesis.

CHAPTER FIVE

CHARACTERISING THE NATURE OF POSITIVE SELECTION

5.1. Introduction

5.1.1. General

A high level of genetic diversity within populations of *T. annulata* has been demonstrated using micro- and mini-satellite loci (Chapters Two and Three). Although such loci are considered to evolve at a high rate, in general there is little selective pressure influencing this variation and mutations may be regarded as neutral (Schlotterer 2000). As previously discussed, this variation principally reflects the differences in motif copy number between alleles and the methodology used for detecting variation takes no account of sequence polymorphism caused by base substitution. Indeed, such polymorphism may explain the subtle variance in designated sequence-length alleles when accurately measured by Genescan™. Since the parasite has been shown to have a primarily panmictic population structure that permits unrestrained and frequent recombination among genotypes (Chapter Three), it may be expected that mutations in an individual have the potential to enter populations and be effectively disseminated. Consequently, it is reasonable to predict a significant amount of sequence diversity between isolates of the parasite from within and between widely separated localities.

Studies in *P. falciparum* have utilised the genetic variation displayed at particular coding regions of the genome to assess the impact of selection on their associated gene products. This parasite is most often found in endemically stable situations with human hosts acquiring a non-sterile protective immunity, which is boosted by repeated exposure throughout life. This is associated with temporal stability of the frequency distributions of alleles at loci encoding antigens (Conway and Polley 2002). Tropical theileriosis exhibits a similar epidemiology in endemic regions, with indigenous cattle developing a non-sterile immunity (see Section 1.6.). Additionally, the bovine stage is obligatory in the life of *T. annulata* resulting in selective pressure from the bovine immune system being applied to every generation of the parasite population. This contrasts with other parasite species such as *Cryptosporidium parvum* (*sensu stricto*), which in humans causes sporadic disease outbreaks (Fayer *et al.* 2000). Since this pathogen is maintained in a reservoir of domestic

and wild animals, it is likely that immune selection from humans plays a minor role in its evolution. However, if one accepts that a similar selective pressure is exerted on *C. parvum* by the immune mechanisms of both animals and humans, a similar pattern genetic variation would be predicted irrespective of host. Due to the obligate bovine stage of *T. annulata*, it may be possible to detect signatures of host immune selection across allelic sequences representing antigen genes. A number of genes were identified in the previous chapter, which exhibited a high interspecies $d_N d_S$ value between a single sequence of both *T. annulata* and *T. parva*. As discussed in Section 4.1.2., positive selection between orthologous genes may be attributed to 'directional selection', with a gene adapting to its own evolutionary niche in each species or 'diversifying selection' as a result of immune selection. Various hypotheses were proposed to explain this evidence of positive selection in the different classes of gene. Putative merozoite surface antigens were identified by bioinformatically screening the genome for genes possessing both a signal peptide and GPI anchor domain and high $d_N d_S$ candidates were proposed to be under the influence of positive 'diversifying' selection from the bovine immune system. Sub-telomeric SVSP family genes were also suggested to diversify, however this may reflect a role in immune evasion. In contrast, TashAT family genes, whose products are postulated to play a critical role in controlling host cell phenotype, were predicted to be under the influence of 'directional' selection. Since *T. annulata* and *T. parva* have differentially adapted to myeloid and lymphoid cells respectively, the genes may have adapted 'directionally' to manipulate the alternate intracellular environments. In order to differentiate between the forces of directional and diversifying selection at different loci, it was necessary to analyse a series of allelic sequences representing isolates of *T. annulata*. Two representative members from each of these three groups of genes were selected for further analysis.

5.1.2. Evidence of selection

The neutral theory of molecular evolution may be used as a null hypothesis to analyse genetic variation. Allelic diversity arises by random mutation and this can be documented by sequence analysis of alleles from populations. Within a species, departure from neutrality may be taken as evidence for selection, whether either positive (diversifying) or negative (purifying). For this reason, sensitive methods of analysing sequence variation have been developed to allow low level or 'weak' selection to be detected in allelic sequences (Wayne and Simonsen 1998). Two of the methods have been used extensively in *P. falciparum* for measuring immune selection on candidate vaccine antigens and will be discussed in detail.

5.1.2.1. McDonald-Kreitman test

McDonald and Kreitman (McDonald and Kreitman 1991) proposed a simple test that examines the number of polymorphic nucleotides which are either synonymous or non-synonymous within a sample of alleles from one species. This is compared with synonymous and non-synonymous polymorphic nucleotides, which are fixed between that species and the orthologous gene of a closely related species. This is an adaptation of the more complex Hudson-Kreitman-Aguade (HKA) test (Hudson *et al.* 1987), which requires sequence data from at least two different loci. The McDonald-Kreitman (MK) test compares patterns of variation within a single gene and is considered more powerful than the Hudson-Kreitman-Aguade test. The MK test is based on the principles of the neutral theory, which may be used to predict the rate of nucleotide substitution over time and the level of polymorphism within a species. If the observed variation in a gene is neutral then the rate of substitution within and between species are both a function of the mutation rate. Therefore the ratio of non-synonymous to synonymous fixed differences between species should be equal to the ratio of non-synonymous to synonymous differences within a species with the ratio, in each case, reflecting the difference in mutation rate of non-synonymous compared to synonymous substitutions. In contrast, if variation within a species is driven by positive selection, an excess of non-synonymous substitutions will be evident among allelic sequences. That is to say, the ratio of non-synonymous to synonymous differences within a species would be greater than the ratio of non-synonymous to synonymous fixed differences between species. Consequently, the null hypothesis of no difference in the number of both types of substitution within and between species can be tested using Fisher's exact test of significance.

This test has been used to identify vaccine targets in *P. falciparum*, using the chimpanzee parasite *P. reichenowi* as the out-species (Escalante *et al.* 1998). Five genes expressed at different stages of the life-cycle encoding candidate antigens were assessed using this method. Two of the five loci show evidence of natural selection – the liver stage antigen, *lsa-1* (Zhu and Hollingdale 1991) and a gene expressed during gametogenesis *pfs48/45* (Kocken *et al.* 1993). In each case the number of amino-acid altering (i.e. non-synonymous) substitutions was greater than expected from the neutral theory. Although, at the time of the study, sequence data was only available for ten loci in *P. reichenowi*, the assumption was made that GC content and codon bias in this species did not affect the MK test. This was justified by measuring the overall GC content, the GC content at the third base and the effective number of codons across these loci and demonstrating that these indices are not statistically different in *P. falciparum*. A potential drawback of the study

was that only a single isolate of *P. reichenowi* was available and it was therefore impossible to assess the level of polymorphism within this species. If, for example, some sites are polymorphic, then the number of fixed differences between the species is likely to be overestimated. However this issue was dismissed because the test compares proportions and there is no reason to speculate that the proportion of non-synonymous to synonymous substitutions will be affected simply by *P. reichenowi* polymorphism. The interpretation of the results, however, with *pfs48/45* was cast into doubt by Conway *et al* (Conway *et al.* 2001) who demonstrated that allelic frequencies at this locus are exceptionally skewed among different *P. falciparum* populations across the world. The parasite isolates, instead of being sampled from a single population, were from different regions of the world. Fixation indices measuring variance in allele frequencies among populations (F_{ST}) were found to be between four and seven times higher at the *pfs48/45* locus than those exhibited by neutral micro-satellite loci. Therefore, the results of the MK test may reflect divergent evolution fixing different alleles in different populations and not the positive / balancing selection sought by the test. It is known that strong directional selection operates in genes involved in mating and this may well be the case in *pfs48/45* as it plays a role in gamete recognition and fertilisation (Palumbi 1999) with the underlying cause of this selection likely to be competition between male gametes (Conway *et al.* 2001).

The MK test has also been successfully applied to identify strong selection on domains of the apical membrane antigen gene, *ama1* (Polley and Conway 2001). 51 *ama1* ectodomain sequences were sampled from a single endemic African population and showed a significant departure from neutrality, with a high excess of intra-specific non-synonymous substitutions. The gene was subdivided into three sub-domains on the basis of predicted structure based on the position of di-sulphide bridges. Similar results were obtained for each domain, providing strong evidence of diversifying selection across the entire gene, although this was only statistically significant for one domain. The test had reduced power in the other two domains as a result of the limited number of fixed differences between the species.

A similar approached may be used in *T. annulata* to assess the impact of positive selection on antigen candidates. *T. parva* would make a suitable choice for the out-species for this test as it is closely related and a complete genomic sequence is available. More importantly, the results presented in Chapter Four suggest that codon usage in *T. parva* is very similar to that in *T. annulata*. Not only is this reflected in the relative synonymous

codon usage (RSCU) values across the genomes, but also the list of putatively optimal codons identified in each species is almost identical. In other words, there is little difference in the codon bias shown by each genome. From the studies in *Plasmodium*, it is clear that the MK test can be of great value in determining genes under positive selection. However, from the *pfs48/45* data in *P. falciparum,* it is also clear that results from this single test taken in isolation may lead to erroneous conclusions, especially when an inappropriate sampling strategy has been used. To avoid this potential problem it has been stressed that more than one method for identifying selection is required before firm conclusions may be drawn (Conway and Polley 2002).

5.1.2.2. $d_N d_S$ ratio

Quantifying the relative rates of synonymous and non-synonymous nucleotide substitutions can determine how nucleotide sequence polymorphism differs from that predicted on the basis of neutrality. This approach has been extensively used in *P. falciparum,* with $d_N d_S$ ratios presented as evidence of positive selection in several candidate antigen genes (Hughes and Hughes 1995; Escalante *et al.* 1998; Verra and Hughes 1999; Anderson and Day 2000). Moreover, using this method supportive evidence has been generated for *lsa-1* and *ama1* genes analysed by the MK test.

In an early study, positive Darwinian selection was identified in four *Plasmodium* genes (Hughes and Hughes 1995). It was suggested that these sporozoite and merozoite surface proteins were under the influence of positive selection exerted by the host immune system, however, this study did not identify positive selection in *lsa-1*. Several candidate genes, including *lsa-1,* produced results consistent with neutral evolution. The method of Nei and Gojobori (Nei and Gojobori 1986) was used to calculate $d_N d_S$, while a previously described method (Hughes *et al.* 1990) was used to classify non-synonymous substitutions as conservative or radical with respect to either amino acid charge or polarity. In the case of the circumsporozoite protein, *CSP*, $d_N > d_S$ occurs in the 3' non-repeat region of the gene where helper and cytotoxic T-cell epitopes have been reported (Good *et al.* 1988; Doolan *et al.* 1991). In the thrombospondin related anonymous protein, *TRAP*, d_N was greater than d_S across all regions of the molecule, with a complex pattern of amino acid substitution. In the relatively conserved N-terminal region, both charge and polarity were conserved, however in the asparagine and proline rich central region there was a greater proportion of substitutions encoding a radically different amino acid (substitutions in the C-terminal region were random). The merozoite surface antigen, *MSA-2*, for which different allelic groups had previously been classified, displayed high intergroup values of both d_N and d_S

in the central region of the gene. This is unsurprising since characteristics of this highly variable region are used to define the groups of alleles. Interestingly, high within-group d_N and d_S values were also observed, with one explanation being a recombination event between groups. Additionally, subgroups of almost identical alleles were observed within which both d_N and d_S were essentially zero, perhaps suggesting the group's recent evolution. In common with the MK test (Escalante *et al.* 1998), the *ama1* gene was identified as being under positive selection where d_N was found to be greater than d_S in the predicted extra-cellular region of the protein. Furthermore, a diversification of amino acid residue charge was noted in this area, while the polarity was conserved. No such effect was seen in the region encoding the signal sequence which is cleaved prior to the protein being mounted on the surface of the merozoite (Kocken *et al.* 2000).

Escalante *et al.* later re-analysed these and other candidate genes (*MSP-1* and *MSP-3*) (Escalante *et al.* 1998) using the same method for calculation of $d_N d_S$ (Nei and Gojobori 1986). Polymorphism was again found to be unevenly distributed among loci, with proteins expressed on the surface of the sporozoite and merozoite found to be more polymorphic than genes encoding intracellular proteins or those expressed during the sexual stages. Seven candidate genes were found to have an excess of non-synonymous substitutions. There was an absence of synonymous changes in three of these loci including *lsa-1* and *pfs48/45*. An alternative method for calculation non-synonymous and synonymous substitution rates (Li 1993) further confirmed evidence for positive selection on six of these seven genes.

The utility of $d_N d_S$ calculated across alleles to identify genes, which are under positive selection in *P. falciparum* is evident. The polymorphic nature of *T. annulata* isolates suggests that a high level of diversity may be found in antigen genes and that $d_N d_S$ analysis of alleles may be feasible. However, it is first necessary to select an appropriately small number of candidate genes for which a sufficient number of alleles can be sequenced.

5.1.3. Novel merozoite antigen candidates

The bioinformatic study described in Chapter Four identified four novel merozoite genes and the already well-characterised *TaMS1* gene as antigen candidates. According to interspecies $d_N d_S$, the two top-ranking candidates were TA16685 and TA08425 (Table 4.1.). These genes are both relatively large at 2.7 kb and 2.9 kb in length, which presented an operational problem since it was necessary to generate a large number of allelic sequences for the McDonald-Kreitman and $d_N d_S$ analyses. Because of the need for

multiple reads to fully sequence each allele, these two genes were not analysed, but the two smaller genes, TA13810 and TA20615 were chosen for further analysis (Table 5.1.) by multiple intra-species allelic sequencing.

TA13810 is the orthologue of a 23 kDa piroplasm antigen identified in *T. sergenti* (Zhuang *et al.* 1995) which is known to exhibit polymorphism between parasite strains (Sako *et al.* 1999). In *T. annulata* it lies in a cluster together with two related genes in the central region of chromosome II (Figure 5.1.(i)). The other two genes have a degree of homology to the *T. sergenti* gene, with protein identity levels of 23 % and 29 % and, although GPI-anchors are present in both, only one has a predicted signal peptide and there is no available expression data from the EST sequences. TA13810 has an identity of 57 % with the *T. sergenti* protein and is therefore the genuine orthologue, although it is predicted to encode a larger protein of 27 kDa. A signal sequence is present with cleavage predicted between amino acid residues 24 and 25. Similar to TaMS1, EST data confirms this protein is expressed in both the merozoite and piroplasm stage and is encoded by a single exon of 690 nucleotides. TA20615 is over twice the size of TA13810 at 1,473 bp and encodes a hypothetical protein product of 57 kDa (Table 5.1.). It is surrounded by spliced genes encoding a diverse range of predicted proteins in the central region of chromosome I (Figure 5.1.(i)). Currently the gene is annotated as a hypothetical protein with EST data indicating that it is expressed in the merozoite. The current gene model has no introns and a signal sequence with a predicted cleavage site between amino acid residues 39 and 40. A transmembrane helix had been predicted at the C-terminus of the predicted proteins of both TA13810 and TA20615, which in each case forms part of a GPI anchor domain. For simplicity, TA13810 and TA20615 will henceforth be referred to as *mero1* and *mero2*. Based on interspecies analysis, there was evidence for selection of these genes and to test whether this was due to diversifying selection imposed by the immune system, a series of alleles were sequenced from populations of *T. annulata*.

5.1.4. Macroschizont gene families

5.1.4.1. SVSP proteins

The SVSP family have been hypothesised to encode variant proteins expressed by the macroschizont of *T. annulata* (Pain *et al.* 2005). Within the genome of the C9 clone of the parasite, members of this family have been identified clustering in all eight sub-telomeric regions. DNA and amino acid sequence alignment of paralogues have indicated a high level of variability among individual family members, outwith the signal peptide sequence.

Table 5.1. Genes selected for allelic sequencing

Six genes were selected for allelic sequencing using DNA from Tunisian and Turkish isolates. Details of the well-characterised merozoite surface antigen, *TaMS1* are also presented as a comparative analysis. '*T. annulata* GeneDB ID' and '*T. parva* ID' correspond to the identification codes used in the GeneDB (www.genedb.org) and TIGR (www.tigr.org) genome databases. d_N, d_S, $d_N d_S$ and nucleotide and protein identities were calculated between the orthologous genes in each species.

Table 5.1. Genes selected for allelic sequencing

Gene name	*T. annulata* GeneDB ID	*T. Parva* ID	d_N	d_S	$d_N d_S$	nucleotide identity (%)	protein identity (%)	*T. annulata* gene						
								Macro	Mero	Piro	Signal	TMD	GPI	NLS
TaMS1	*TA17050*	*TP01_1056*	*0.198*	*0.720*	*0.2751*	*72.86*	*77.38*		√	√	√		√	
mero1	TA13810	TP02_0551	0.097	0.592	0.1638	84.57	83.41		√	√	√	1	√	
mero2	TA20615	TP01_0487	0.199	0.776	0.2560	76.92	70.59		√		√	1	√	
SVSP1	TA16025	TP02_0955	0.467	1.157	0.4037	65.60	43.21	√			√			
SVSP2	TA17485	TP01_1225	0.477	1.530	0.3118	62.40	44.97	√	√		√			
TashHN	TA20090	TP01_0603	0.218	0.822	0.2646	77.11	65.36	√ *			√			√
SuAT$_1$	TA03135	TP01_0617	0.449	2.048	0.2192	62.93	47.40	√ *			√			√

d_N = rate of non-synonymous substitutions, d_S = rate of synonymous substitutions, $d_N d_S$ = non-synonymous to synonymous substitution rate

Macro, Mero & Piro = macroschizont, merozoite and piroplasm EST data, Signal = signal peptide motif, TMD = transmembrane domain,

GPI = glycosyl-phosphatidylinositol anchor, NLS = nuclear localisation signal

* experimental data

Figure 5.1. Schematic diagram of genes encoding secreted products

The location of the merozoite, SVSP and host-nuclear genes under study are marked on schematic diagrams representing chromosomes of *T. annulata*. For each locus, an increased resolution diagram of the chromosome displays the arrangement of genes neighbouring the gene of interest; the position of coloured blocks either above or below a central line indicates the coding strand of DNA. For each of the six genes under study, a schematic diagram represents the translated gene product with peptide motifs indicated.

(i) Merozoite genes (*mero1* and *mero2*)

Both merozoite genes are located in internal regions of chromosomes I and II. *mero1* is flanked by two related genes.

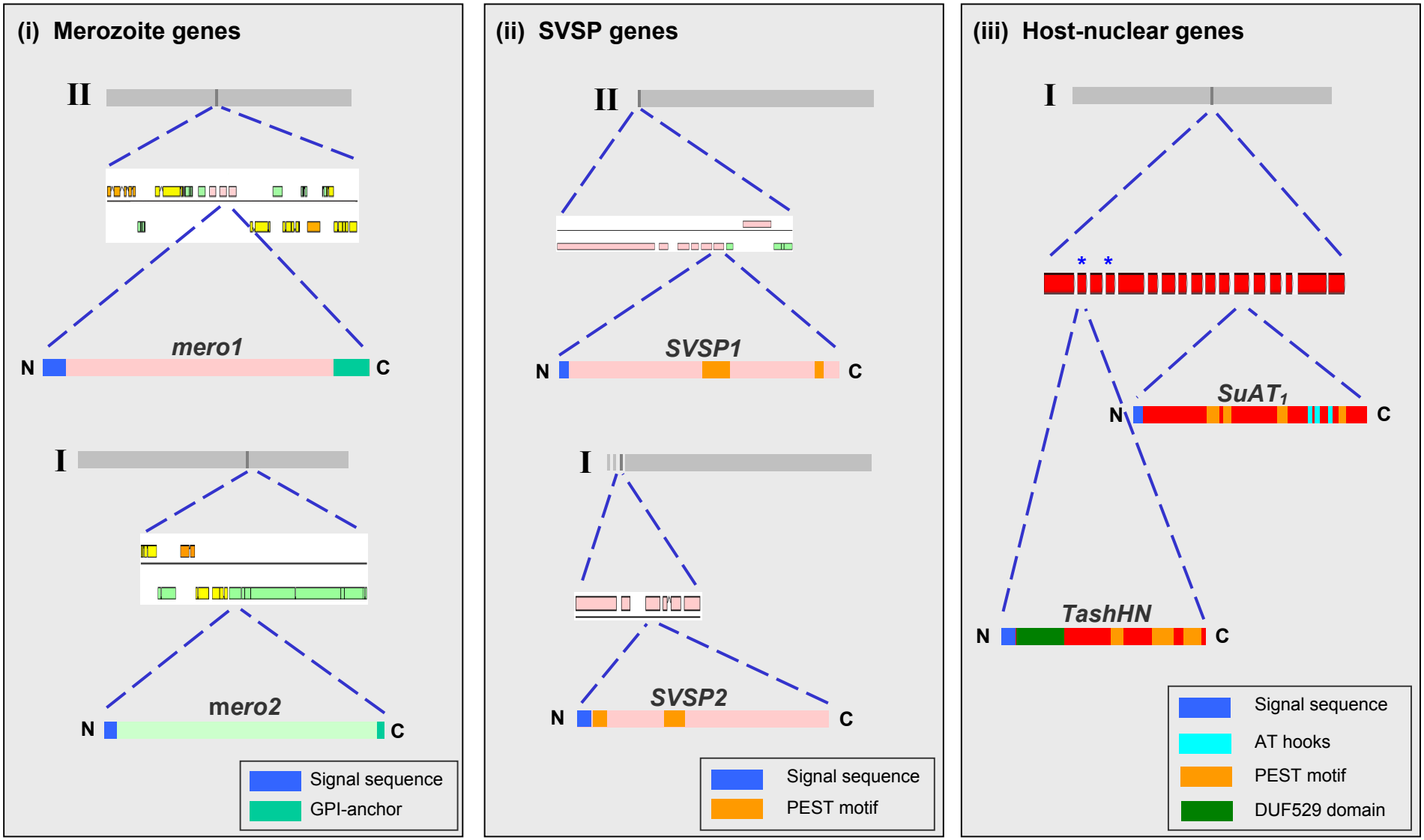
(ii) SVSP genes (*SVSP1* and *SVSP2*)

SVSP family genes are located in tandem arrays at each of the eight sub-telomeric regions. *SVSP2* is located on one of three short contigs that have not yet been incorporated into the model of chromosome I.

(iii) Host-nuclear genes (*TashHN* and $SuAT_1$)

The TashAT family is located internally in chromosome I, comprising 16 distinct genes encoded on the same strand. In the *T. annulata* (Ankara C9) genome, two copies of *TashHN* are present and are denoted by asterisks (*) on the diagram.

Figure 5.1. Schematic diagram of genes encoding secreted products



Sub-telomeric regions are generally considered to have atypical properties, including reversible silencing of genes mediated by protein binding to telomeres and intragenic recombination between sub-telomeres (Barry *et al.* 2003). Such ectopic recombination has been demonstrated in *S. cerevisiae* (Britten 1998) and is understood to be promoted through the extensive sharing of sequence elements in these regions coupled with the fact that telomeres cluster at the nuclear periphery. In contrast, sub-telomeric regions in yeast have also been subject to a degree of homogenisation not observed in other regions of the genome, which has been explained by several recombination mechanisms (Louis and Haber 1990).

The identification of this sub-telomeric gene family with each member having a signal sequence suggesting surface expression raises the question of whether there is a system of antigenic variation in *T. annulata* analogous to that found in *P. falciparum* and African trypanosomes. These highly variable sets of genes involved in immune evasion are well documented and have been termed contingency genes (Moxon *et al.* 1994). The case of the SVSP genes may be most analogous to the mechanisms employed by *Plasmodium*, where contingency genes are expressed by an intra-cellular stage. *P. falciparum* undergoes antigenic variation with at least two classes of proteins being expressed on the surface of the erythrocyte (Kyes *et al.* 2001). Again these are encoded by large gene families, *var* and *rif*, which display differential expression. There are around 60 paralogues of the *var* gene present in the genome varying in size from 4 to 13 kb, which are distributed over all the chromosomes with most members located in sub-telomeric arrays (Rasti *et al.* 2004). At least one paralogue is present in the sub-telomeric regions of each chromosome. The *var* genes in the sub-telomeres are known to be more vulnerable to recombination compared to the more conserved centrally located genes (Rasti *et al.* 2004). The largest family implicated in *P. falciparum* antigenic variation is the *rif* genes (repetitive interspersed family), with 149 members found in the 3D7 genome (Gardner *et al.* 2002). Similar to *var* genes, *rif* genes are largely encoded at the sub-telomeres and like the products of the *var* genes are translocated to the surface of the erythrocyte. In the genome, *rif* genes are always adjacent to *var* genes implicating some relatedness in their functions. A third, smaller family of variant sub-telomeric genes, *stevor* (sub-telomeric variable open reading frame) have also been uniquely identified in *P. falciparum*. These genes are only found adjacent to *rif* genes and are believed to be more conserved among different parasite genotypes (Blythe *et al.* 2004). The products of *stevor* genes are located in Maurer's Clefts in the sub-membrane of the infected erythrocyte (Kaviratne *et al.* 2002). Only subsets of these gene families are expressed simultaneously, and *in vitro* experiments have

demonstrated that the rate of switching away from a parental variant type was 2 % per generation (Roberts *et al.* 1992). A variant antigen theory may predict cyclical patterns of expression across the family, with only a subset of genes being expressed at any one time. The bloodstream form of African trypanosomes employ a strategy of sequentially varying antigenic types of the variant surface glycoprotein (VSG), which physically surrounds the parasite. This protective coat undergoes antigenic variation, spontaneously producing waves of variant protein to stimulate an immune response by the host (Barry and McCulloch 2001). VSG genes are located in arrays, some of which are at sub-telomeric loci. The mechanisms controlling expression are complex, however they permit only a single gene expression site to be active at any one time, leaving the majority of genes unexpressed. However, four important pieces of evidence suggest that SVSP genes are not involved in a classical system of antigenic variation on the parasite or host cell surface - (1) there are no genomic signatures such as GPI-anchors or TMDs to suggest that SVSP proteins are membrane bound to either the surface of the macroschizont or the infected leucocyte; (2) previous studies have failed to detect any parasite encoded proteins on the leucocyte surface (Preston *et al.* 1999); (3) as previously discussed, SVSP proteins contain PEST domains which may promote their rapid degradation and (4) current EST data indicated that the majority of SVSP members were all expressed in a single culture of the C9 clone (Pain *et al.* 2005). This would suggest that a classical antigenic variation system is unlikely, due to an apparent lack of sequential variation. However, it should be considered that the EST data may simply reflect differential expression within the population of cultured cells, with individual cells expressing different paralogues. Taken together, these facts suggest that SVSP proteins are secreted into the host cell, degraded and are therefore liable to enter a processing pathway for presentation on the surface of the infected cell via MHC molecules. If they are presented in this way and are recognised by the protective immune response, one would predict that the products would be subjected to diversifying selection. However, this does not explain why there are so many diverse family members that appear to be constitutively expressed.

The genes TA16025 (*SVSP1*) and TA17485 (*SVSP2*) were chosen as they exhibit features typical of SVSP proteins, including macroschizont expression, possession of a signal sequence, a PQ rich region and multiple PEST motifs. Details of these two members along with their chromosomal location are presented in Table 5.1. and Figure 5.1.(ii). *SVSP1* is located in a sub-telomeric region of chromosome II, whereas *SVSP2* is found in one of the three short contigs predicted to encode sub-telomeric regions of chromosome I. In common with two thirds of putative SVSPs, these genes have orthologues identified in

T. parva and possess relatively high interspecies $d_N d_S$ values of 0.4037 and 0.3118 respectively. Similar to the merozoite antigen candidates, the length of each gene was also taken into consideration when identifying SVSPs suitable for cloning and sequencing. The current model predicts that the gene products are encoded in each case by a single exon of 1,683 bp and 1,233 bp.

5.1.4.2. Parasite encoded host-nuclear proteins

Merozoite candidate antigens and SVSP family members represent genes, which lie toward one extreme of the spectrum of expected diversity, where a high degree of polymorphism is predicted. To provide a context for interpreting the results for merozoite and SVSP genes and to investigate the potential that a proportion of the *Theileria* genome may be under directional selection, it was decided to include an additional set of genes in the analysis of allelic diversity.

The TashAT family was identified in Chapter Four as having a high mean $d_N d_S$ value across members. This cluster of 16 related genes is found towards the centre of chromosome I and is depicted in Figure 5.1.(iii). The expressed proteins of several members of this family have been shown to locate to the host cell nucleus as well as bearing AT hooks, a form of DNA-binding motif (Swan *et al.* 1999; Swan *et al.* 2001) and it has been proposed that this family of genes is involved in altering bovine gene expression and manipulating leucocyte phenotype. BLAST search results have identified the TashAT family as part of a larger family including the SVSP family and other sub-telomeric genes. In common with SVSP genes, parasite-encoded host nuclear proteins generally possess a signal sequence, multiple PEST sequences and have an abundance of glutamic acid residues. Multiple TMDs are absent however a single transmembrane helix is identified in some members of the family as part of the signal sequence. In contrast to the putative role of SVSPs, these proteins are believed to perform specific activities since they have been shown to contain conserved functional domains (Swan *et al.* 1999; Swan *et al.* 2001). Nuclear localisation signals (NLS) have been identified bioinformatically in 8 of the 16 gene family members, a subset of which contain DNA-binding motifs, which include the characterised AT-hook domains. Additionally, several family members contain the novel DUF529 domain (domain of unknown function) (Pain *et al.* 2005). This domain represents a repeated region also found in several *T. parva* proteins. The repeat is normally about 70 residues long and contains a conserved central aromatic residue. Since TashAT family proteins are secreted into the host cell compartment before locating to the host nucleus and since they possess PEST motifs, it may be predicted that they enter a

similar pathway to that proposed for SVSP proteins, i.e. proteolysis and presentation by host MHC molecules. Consequently this gene family may be predicted to exhibit evidence of diversifying immune selection. However, considering the evidence of functionality, a significant level of amino acid sequence conservation may be expected at this locus within the genome of *T. annulata*. Hence, a balance between immune selection and purifying selection may be predicted among alleles of the genes comprising the TashAT cluster.

Two members of this family were selected for allelic sequencing. *TashHN* (TA20090) is a tandemly duplicated gene in the C9 strain of the parasite whose product is known to be transported to the host nucleus during the macroschizont stage of infection (Swan *et al.* 2003). *SuAT1* (TA03135) is a single-copy gene and like *TashHN* is predicted to comprise a single exon. Functional domains in this protein are located towards the C-terminus, as opposed to across the entire molecule in *TashHN* (Figure 5.1.(iii)). Orthologues of both genes have been identified in *T. parva* and high interspecies $d_N d_S$ values of 0.2646 (*TashHN*) and 0.2192 (*SuAT1*) were observed (Table 5.1.). The genes are an appropriate size for cloning and sequencing at 999 bp and 1,677 bp respectively.

5.1.5. Sampling rationale

With a panel of six genes selected to test for evidence of positive selection, a suitable sampling strategy must be defined to generate sets of sequences that will accurately inform on the nature of the selective pressure at each locus.

Selection of advantageous genotypes is a local phenomenon with allele-specific immune responses promoting the survival of unrecognised individuals. As future generations of the novel genotype expand, a growing proportion of the host population becomes immune to many alleles. Since there is generally limited transit of cattle and vectors over large geographical distances in endemic regions, it is unlikely that widely separated populations of *T. annulata* will be exposed to the same bovine immunological background. Consequently, in order to confidently relate sequence diversity to the effect of balancing selection, one must sample from a single parasite population. The population genetic studies presented in Chapters Two and Three suggest there is a degree of sub-structuring of the parasite between the countries of Tunisia and Turkey, although limited genetic differentiation is evident within these regions. Geographical sub-structuring is not, however, reflected in the distribution of *TaMS1* alleles (Gubbels *et al.* 2000b). For this antigen, a high level of sequence diversity was observed both within localities and across different parts of the world. A drawback of the latter study was that large samples from

within discrete geographical regions were not compared, so the results must be interpreted with some caution. Because the level of diversity that would be encountered for each of the genes selected for analysis was difficult to predict, a two-pronged sampling strategy was adopted to maximise the variation, but also to represent individuals from genuinely sympatric populations.

1. Sampling within a very limited geographical range. This approach involves a sampling strategy concentrated on a single farm or village. The sequence of every allele that is generated would originate from parasites within a small number of cattle within close proximity of one another. The cattle would be managed similarly and be exposed to an identical population of ticks. Therefore, one may have confidence that the genotypes under study were subjected to similar selective pressures. However, the level of diversity encountered may be limited with identical and closely related alleles being found across the sample. This may result in a limited amount of sequence polymorphism, which could prove insufficient for the statistical tests proposed and for this reason an additional strategy was considered.

2. Sampling over a moderately large geographical range. The micro- and mini-satellite data indicated little genetic differentiation throughout Tunisia and it may therefore be regarded as a single population for genetic studies. In order to maximise the chance of detecting polymorphism in candidate genes a second strategy was undertaken to encompass several different locations across the country. Allelic sequence diversity calculated across wide regions may still reflect positive or purifying selection generated at a local level. That is to say, the fact that immune selection operates locally does necessarily mean that it cannot be detected over large geographical areas.

5.2. Materials and methods

5.2.1. Parasite material

Micro- and mini-satellite genotyping of *T. annulata* has shown a very high level of heterogeneity in field isolates with mixed infections being almost universal. A particularly high multiplicity of infection was noted in cattle from the village of Sariköy in the Aydın region of Western Turkey. DNA preparations from the blood of two indigenous and two dairy-type unvaccinated male cattle between 10 and 21 months old were selected for sequence analysis. These blood samples were taken from the cattle in late 2001 and are described in Table 5.2.

Table 5.2. Turkish isolates used for sequencing

Micro- and mini-satellite genotyping identified four highly heterogeneous samples, which were isolated from unvaccinated male cattle in the village of Sariköy, Akçaova district in the Aydın region of western Turkey.

Table 5.3. Tunisian clones used for sequencing

Micro- and mini-satellite analysis identified ten Tunisian DNA preparations containing different clonal genotypes.

Table 5.2. Turkish isolates used for sequencing

Study ID	Sample date	Origin	Animal Owner	Breed	Age (months)	MS genotyping (alleles / locus)	
						Mean	Max
t005	September 2001	piroplasm	Bahtiyar Demirel	Indigenous	16	5.4	9
t021	December 2001	piroplasm	Nuri Katidemirel	Brown Swiss	10	5.6	11
t029	December 2001	piroplasm	Bahtiyar Demirel	Holstein	18	5.7	10
t038	December 2001	piroplasm	Bahtiyar Demirel	Indigenous	21	6.3	12

MS = micro- and mini-satellite

Table 5.3. Tunisian clones used for sequencing

Study ID	Additional information	Origin	Passage number	Bioclimatic zone *	Site *
w019	14B *	cell line	p10	Semi arid (C)	1
w027	19 *	cell line	p9	Semi arid (C)	4
w030	22 *	cell line	p18	Semi arid (C)	4
w032	24A (2c) *	cell line	unknown	Semi arid (C)	3
w050	42 *	cell line	p8	Semi arid (C)	13
w062	527 cl 4 *	clone	-	-	-
w067	29 cl 4 *	clone	-	-	-
w073	Clone 2 BV4 Cl5 *	clone	-	-	-
w101	Jedeida 4	cell line	p200	Semi arid (C)	-
w102	Batan 2	cell line	p198	Semi arid (C)	-

* Details in PhD thesis of Leila Ben Miled

A small number of parasite preparations were identified in Tunisia that contained a single haploid genotype, with a single allele at each locus. Using a sample set such as this for generating allelic sequences would be advantageous since candidate gene sequences can be related to the micro- and mini-satellite MLG. A panel of these preparations (n = 10), each having a different MLG, was created from the Tunisian DNA collection held at Glasgow University Veterinary School and is detailed in Table 5.3. Five of these isolates are derived directly from cell lines established by sampling cattle from different sites in the semi-arid region of Tunisia (Ben Miled *et al.* 1994). Two DNA preparations were derived from the Tunisian cell lines Jedeida4 and Batan2. The absence of heterogeneity in these lines may be attributed to the fact they each had been passaged around 200 times. Three further isolates in the form of clonal cell lines were also included in this set (Ben Miled 1993).

A DNA preparation from the genome clone (C9) was also included in the study as a positive control for the PCR primers used to amplify alleles and also to determine the sequencing error rate.

5.2.2. PCR amplification of alleles

Forward and reverse PCR primers for each of the six genes were designed in the signal peptide and 3' downstream sequences respectively, with only two exceptions: (i) the forward primer for *SVSP1* was located in the 5' region, slightly upstream of the coding sequence and (ii) an intragenic reverse primer was designed for $SuAT_1$ for use in samples that would not amplify with the primer located in the 3' sequence. The design was based on the published genome sequence (Pain *et al.* 2005) and the oligonucleotide sequences are detailed in Table 5.4.

An aliquot of each DNA preparation was PCR amplified in a total reaction volume of 20 µl under conditions previously described (MacLeod *et al.* 2000), using a Techne TC-512 thermocycler with the following settings: 94 ºC for 2 minutes, 30 cycles of 94 ºC for 50 seconds, 50 ºC for 50 seconds and 65 ºC for 90 seconds with a final extension period of 15 minutes at 65 ºC. A mixture of *Taq* polymerase and a proofreading polymerase (*Pfu*) at a ratio of 15:1 was used to improve the fidelity of the reaction, while maintaining the 3' A-overhangs on the PCR product that are required for TOPO® cloning. Amplicons were separated by electrophoresis on a 1.5 % agarose gel and stained with ethidium bromide solution. Gels were photographed under ultra-violet transillumination and the size of each

Table 5.4. PCR and sequencing primers

Forward and reverse PCR primers were designed to amplify the six genes based on the published genome sequence of *T. annulata* (Pain *et al.* 2005). The numbers in parenthesis refer to the nucleotide position with reference to the C9 genome clone. Two different reverse primers were required to amplify *SuAT*$_1$ from the Tunisian and Turkish DNA samples (Figure 5.2.). Conserved internal primer sites for sequencing were identified following alignment of initial sequencing reads generated by the vector primer sites (M13).

Table 5.4. PCR and sequencing primers

Gene name	GeneDB ID	PCR primers		Sequencing primers (sense & anti-sense)	
		Forward	Reverse	Internal locus 1	Internal locus 2
mero1	TA13810	TGTCCTCTTGACACACGC (30-47)	CGTTAGTGTGTGAGATCGAGG (3' UTR)	-	-
mero2	TA20615	GCGGGAAGAGAGAAGTGTTG (8-27)	GTATGTAACTAGATCCCCATG (3' UTR)	GAGTTGAGGTCAATAGGG CCCTATTGACCTCAACTC	AGTAGTGGTTACCTGGGT ACCCAGGTAACCACTACT
SVSP1	TA16025	CATGGGTCAATGTCAAATAACATTC (5' UTR)	CATAAACTTACATCATATAG (3' UTR)	CAGACCCTCCTTTACCTA TAGGTAAAGGAGGGTCTG	CAGTGTTTCATTGTCTACTAGG CCTAGTAGACAATGAAACACTG
SVSP2	TA17485	ATGAATAAATACGTTAGATACAC (1-23)	GACGATTCTAAGTTTTATGTGC (3' UTR)	CCTCCAGCAATTGAGTAT ATACTCAATTGCTGGAGG	-
TashHN	TA20090	ATGACCAGATTAAAGATTGC (1-20)	GTGTTCAATTATGGTGGCTTGTG (3' UTR)	GACCAACTCATTCAAGAG CTCTTGAATGAGTTGGTC	-
SuAT$_1$	TA03135	CCTTGTATTGTTTTTACAG (19-37)	GATGATTTGTTTCATGTCTTC (3' UTR) * GGTTGTAATTCTAAATGTTCTGG (1487-1465) †	TGAAGAGACGGATACTGC GCAGTATCCGTCTCTTCA	CCTAGGATACGTAGACCT AGGTCTACGTATCCTAGG

* used to amplify Tunisian clones

† used to amplify Turkish isolates

PCR product determined with reference to a 100 bp DNA ladder (Promega). These PCR products were stored at -20 ºC for subsequent cloning reactions.

5.2.3. Cloning and sequencing

PCR products were cloned into the sequencing vector pCR4®-TOPO® (Invitrogen). This vector is specifically designed to clone *Taq* polymerase-generated PCR products for sequencing and has the advantage of containing commonly used sequence-priming sites closely flanking the site of insertion, thereby limiting the amount of vector sequence generated. Ligation of PCR product disrupts expression of the lethal *E. coli* gene, *ccd*B, permitting selection of only recombinants upon transformation into competent cells. Cells that contain non-recombinant vector are killed upon plating using selective media containing an antibiotic. The cloning reaction and transformation procedure were carried out as directed in the 'TOPO TA Cloning® Kit for Sequencing' user manual. To summarise, for each reaction, an aliquot of PCR product was incubated at room temperature for five minutes together with the vector and then transformed into TOP10 *E. coli* chemically competent cells, by heat shocking at 42 ºC. This reaction was incubated for one hour at 37 ºC before being spread on Luria Broth (LB) agar plates containing a selective agent - either ampicillin or kanamycin. Plates were incubated overnight, after which bacterial colonies were picked and used to seed a culture of 8 mls of LB. These cultures were grown for 16 hours at 37 ºC before the bacterial cells were harvested. Up to 20 µg of plasmid DNA was isolated from each culture with a proprietary kit (Qiagen) using the technique described in the QIAprep® Miniprep Handbook. After the quantity and quality of DNA was determined by spectrophotometer, 1 – 2 µg of air dried DNA was prepared for use in each sequencing reaction, which was performed using a commercial sequencing service (MWG Biotech, Germany).

M13 universal (forward) and M13 reverse primer sites, present in the vector flanking sequence, were used in the initial sequencing reactions to confirm the presence of an appropriate insert. This process allowed complete insert coverage in both directions for the smallest gene (*mero1*). Internal priming sites were identified for the other five genes, based on conserved sequence generated from the initial sequencing reads using the M13 primers. Sense and anti-sense oligos were designed corresponding to a single internal site for *TashHN* and *SVSP2* and two sites in *mero2*, *SVSP1* and $SuAT_1$, to allow at least two times coverage of the insert. These sequencing primers are also detailed in Table 5.4.

Sequence data was received electronically in Standard Chromatogram Format (SCF). Corresponding FASTA files were generated and used to assemble complete consensus sequences with the ContigExpress feature of the software package Vector NTI Advance™ (Invitrogen). Wherever present, ambiguous nucleotide sequence was re-examined in SCF files before the decision was made to accept or re-analyse the sequence.

5.2.4. Analytical tools

5.2.4.1. Cluster analysis

ClustalX (Thompson *et al.* 1997) is the current Windows version of the Clustal software program (Higgins and Sharp 1988), which is commonly used for aligning both nucleotide and amino acid sequences. It operates by initially aligning closely related sequences following which, groups of sequences are progressively incorporated into a complete alignment. PHYLIP trees were generated from the aligned sequences and were viewed using TreeViewX (Page 1996).

5.2.4.2. The McDonald-Kreitman test

DNA sequence polymorphism was evaluated using the computer package DnaSP (Rozas and Rozas 1999; Rozas *et al.* 2003). The software was used to estimate several measures of DNA sequence variation within and between populations including the McDonald-Kreitman test (McDonald and Kreitman 1991). To test whether the level of synonymous or non-synonymous polymorphisms deviated from the neutral prediction of equal numbers either within *T. annulata* or between species, Fisher's exact test of significance was applied to the 2 x 2 matrix containing the results for each gene; a low *p* value reflected a departure from neutrality. The 'neutrality index' odds ratio was also calculated for each locus (Rand and Kann 1996). This measurement was used to indicate if there was an excess (ratio > 1) or deficiency (ratio < 1) of non-synonymous substitutions within *T. annulata* alleles and was used a qualitative and quantitative indicator of the direction and degree of selection.

5.2.4.3. $d_N d_S$ analysis

A maximum likelihood (ML) method was used to detect amino acid sites under positive selection and to determine $d_N d_S$ values across the alleles (Pond and Frost 2005b). This was undertaken using the molecular evolution analysis platform HyPhy (Pond *et al.* 2005) via the Datamonkey web interface (Pond and Frost 2005a) at the URL http://www.datamonkey.org/. This heavily modified version of the Suzuki-Gojobori

method (Suzuki and Gojobori 1999) is termed Single Likelihood Ancestor Counting (SLAC). SLAC tends to be more conservative than the alternative random effects likelihood (REL) and fixed effects likelihood (FEL) methods and may be used to analyse up to 150 sequences. A four-phase process is used to analyse the aligned sequences and identify sites under positive or purifying selection and then calculate the significance of this observation- (1) a neighbour joining tree is first created and a nucleotide model is fitted to both this tree and the dataset using a maximum likelihood approach, which determines tree branch lengths and substitution rates; (2) keeping branch lengths and substitution rate parameters constant, a codon model is generated and a global $d_N d_S$ ratio is obtained; (3) using the optimum nucleotide and codon models, ancestral sequences are reconstructed on a site by site basis using maximum likelihood and (4) for every polymorphic codon, expected and observed numbers of synonymous and non-synonymous substitutions are calculated (ES, EN, NS, NN). This permits the $d_N d_S$ ratio to be determined for each codon where d_N = NN / EN, d_S = NS / ES and if d_N is greater or less than d_S a *p* value derived from a two-tailed binomial distribution is calculated to assess the significance.

5.2.4.4. Neutrality tests

Tajima *D* statistics (Tajima 1989) were calculated to evaluate the allelic distribution of the genes selected for this study. This test was developed to determine the neutrality of mutations by comparing two estimators of genetic variation (θ). The first, θ_π is an estimate of variation calculated by examining nucleotide diversity (π) ($\theta_\pi = 4N_e\mu$, where N_e is the effective population size and μ is the mutation rate). This is compared with a second estimate θ_S, which is based on the number of segregating sites (S). Although these estimates reflect different types of information, they are expected to be of equal value under the neutral theory (Kimura 1969). The test compares the difference between these two estimates ($d = \theta_\pi - \theta_S$) with the *D* value being computed by dividing d by the standard deviation of d. If a population is at neutral equilibrium, Tajima's *D* value should be zero. If some mutations are slightly deleterious then π would not be influenced significantly, but S would increase. Consequently θ_π would be smaller than θ_S and the difference would be negative. A negative Tajima *D* value indicates an excess of low frequency alleles relative to neutral mutation–drift equilibrium. A positive Tajima *D* value indicates a deficit of low frequency alleles relative to expectations. This could be due to a population bottleneck, population subdivision or balancing selection. However, the power of the test may be limited to detecting either strong or recent rounds of positive selection (Simonsen *et al.* 1995).

Fu and Li's *D* test of neutrality was also utilised (Fu and Li 1993). This similar, but potentially more informative test is based on a comparison of the number of mutations in the internal and external branches of inferred phylogenetic trees and measures recent versus ancient mutations. Balancing selection may be indicated by an increased number of internal branches, whereas an excess of external branches would imply purifying selection (Hedrick 2005). The value is based on the differences between the number of singletons (mutations appearing only once among the sequences) and the total number of mutations (Fu and Li 1993). Also used was the related Fu and Li's *F* test statistic (Fu and Li 1993), which is based on the differences between the number of singletons and the average number of nucleotide differences between pairs of sequences.

Fu and Li's *D* and *F* tests and Tajima's *D* test were performed using the computer package DnaSP (Rozas and Rozas 1999; Rozas *et al.* 2003). The software was also used to estimate the confidence intervals of these neutrality test statistics by coalescence modelling.

5.3. Results

5.3.1. Sequencing data

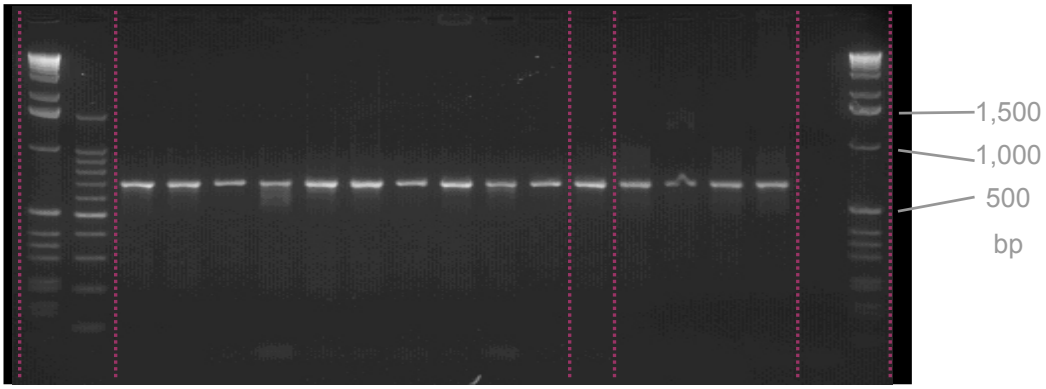
PCR products were successfully generated for the majority of primer and template combinations for all six genes. They were checked for insert size by electrophoresis on an agarose gel before being cloned into the sequencing vector. Examples of images from agarose gel electrophoresis are presented in Figure 5.2., representing products of amplification using sets of primers corresponding to *mero1* and $SuAT_1$. The *mero1* primers (Figure 5.2.(i)) can be seen to generate a PCR product of similar size using each of the Turkish (Table 5.2.) and Tunisian (Table 5.3.) isolates as a template for amplification. For $SuAT_1$ two separate sets of primers were required to amplify from Tunisian and Turkish isolates, corresponding to a combination of a forward primer designed in the sequence encoding the signal peptide along with two different reverse primers. The reverse primers were located (i) in the non-coding 3' UTR and (ii) intragenically, around 200 bp from the 3' end of the coding sequence. As expected, both primer combinations were able to amplify C9 DNA (Figure 5.2.(ii) and (iii)), but no insert-positive colonies were subsequently generated. For each gene / template combination, PCR products were cloned into a sequencing vector as described in Section 5.2.3. For each of the four Turkish isolate DNA templates, 24 plasmid colonies corresponding to *mero1* PCR product were grown and sequenced, i.e. 4 x 24 colonies. For each of the other five genes, eight different colonies were grown and sequenced representing each of the four Turkish isolates, i.e. 5 x 8 x 4

Figure 5.2. Amplification of *mero1* and *SuAT*$_1$

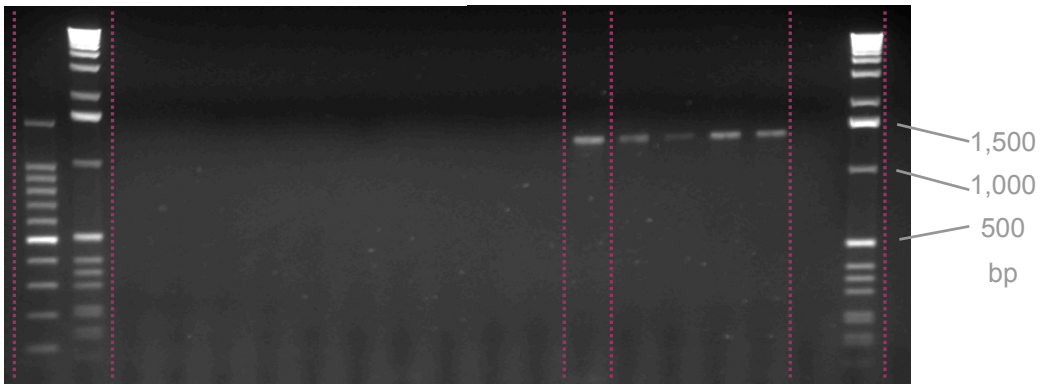
DNA representing the panel of ten Tunisian clones and four heterogeneous Turkish isolates (Tables 5.2. and 5.3.) was used as template for PCR amplification of six loci. PCR products were separated by agarose gel electrophoresis, examples of which are presented opposite. Two alternate sets of PCR primers were used to amplify from Tunisian and Turkish isolates at the *SuAT*$_1$ locus.

Figure 5.2. Amplification of *mero1* and *SuAT$_1$*

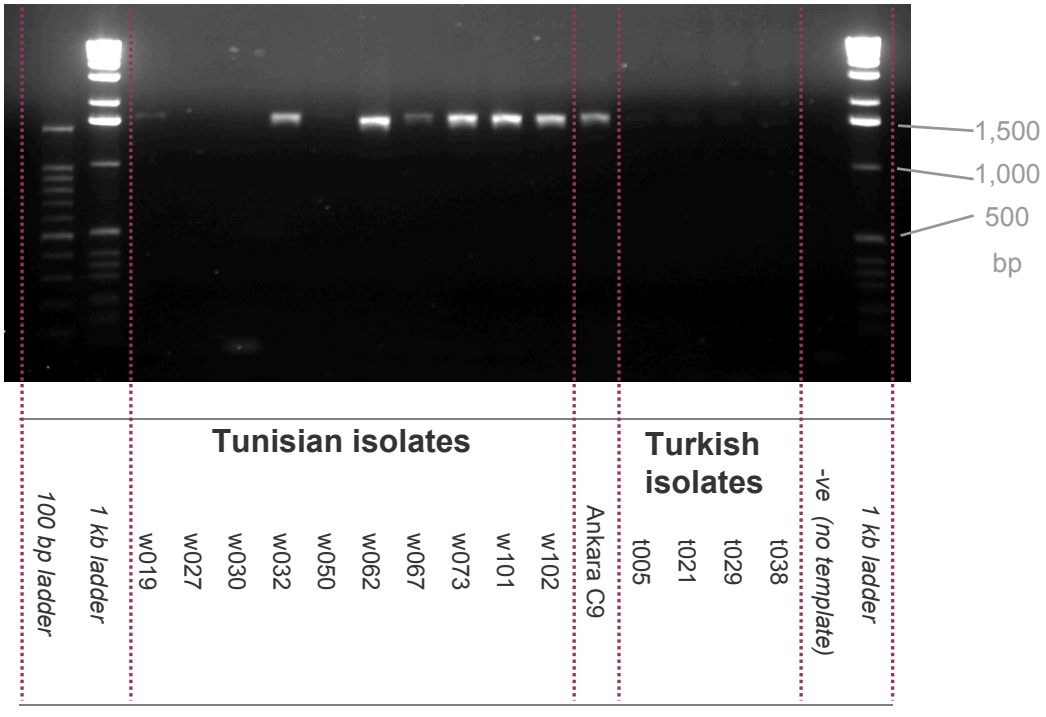
(i) *mero1*



(ii) *SuAT$_1$* (Primer set one)



(iii) *SuAT$_1$* (Primer set two)



colonies. For the ten Tunisian DNA preparations, two clones were fully sequenced for the *mero1* PCR product and for the other five genes one clone was fully sequenced and a second partially sequenced to check for polymorphism. All of the sequence data from each of the plasmid preparations was assembled using Vector NTI to provide the complete sequences for the majority of preparations. Consistent with the earlier micro- and mini-satellite genotyping, comparison of these sequences revealed each Turkish isolate contained a number of different alleles at each locus, whereas only a single allele at each locus was identified from each of the Tunisian preparations. A summary of the number of unique alleles identified at each locus in each DNA template is presented in Table 5.5. The mean number of unique alleles at each locus was calculated, although *mero1* was not included in this calculation because sequences were derived from 24 rather than 8 colonies and this would have distorted the mean value. For the Turkish sample, the mean value ranged from 2.4 for t029 to 3.6 for t005 and gave an indication of the amount of diversity identified within each of the four cattle blood preparations. Presented in a slightly different way, the proportion of unique alleles in the Turkish isolates that each DNA sample represented was calculated and is shown in Figure 5.3. Isolate t005 can be seen to represent a consistently high proportion of unique alleles for each of the six genes, ranging from 21 % of *mero1* alleles to 62 % of $SuAT_1$ alleles. Between 12 % and 38 % of alleles for each gene were identified using t038 as a template, whilst a larger variance was displayed by both t021 and t029.

For all genes, a number of unique sequences were identified within both the Tunisian and Turkish isolates. 47 alleles of *mero1* were identified across the Turkish collection of 96 clones, therefore almost every second clone analysed represented a novel sequence. Between 6 and 13 alleles were identified for the other genes in the study across the Turkish isolates. The Turkish samples PCR-amplified particularly well, with products of an expected size generated from all four DNA preparations for all of the six genes under study. No sequences of the *mero2* gene were obtained using the t029 sample as none of the colonies picked and grown contained an insert. Although in theory plasmids without an insert could not grow, in practice a proportion of such colonies did survive. Similarly, only four of the ten Tunisian DNA templates were successfully cloned and sequenced for this gene. This is surprising, since *mero2* PCR-amplified well across the panel of Tunisian samples (data not shown). The low number of inserts obtained is probably due to a low efficiency of transformation. A greater proportion of insert-positive colonies were obtained for the other five genes ranging from seven (*SVSP2*) to ten (*mero1* and *SVSP1*) in the Tunisian samples.

Table 5.5. Summary of unique sequences derived from Tunisia and Turkey

The Turkish samples represented mixed genotypes while the Tunisian samples represented clones, as predicted from micro- and mini-satellite genotyping. The number of *mero1* alleles identified was not included in the mean figure for each Turkish isolate since 24 colonies of *mero1* were sequenced for each isolate while only 8 colonies were sequenced for each isolate for the other five genes. Due to identical alleles being identified in different isolates, the subtotals of unique alleles within each country do not necessarily equal the sum of the unique alleles identified in each isolate from that country. Similarly, for some genes, the sum of the subtotals for each country is greater than the total number of unique alleles (on the bottom line), due to the presence of identical alleles in Tunisia and Turkey.

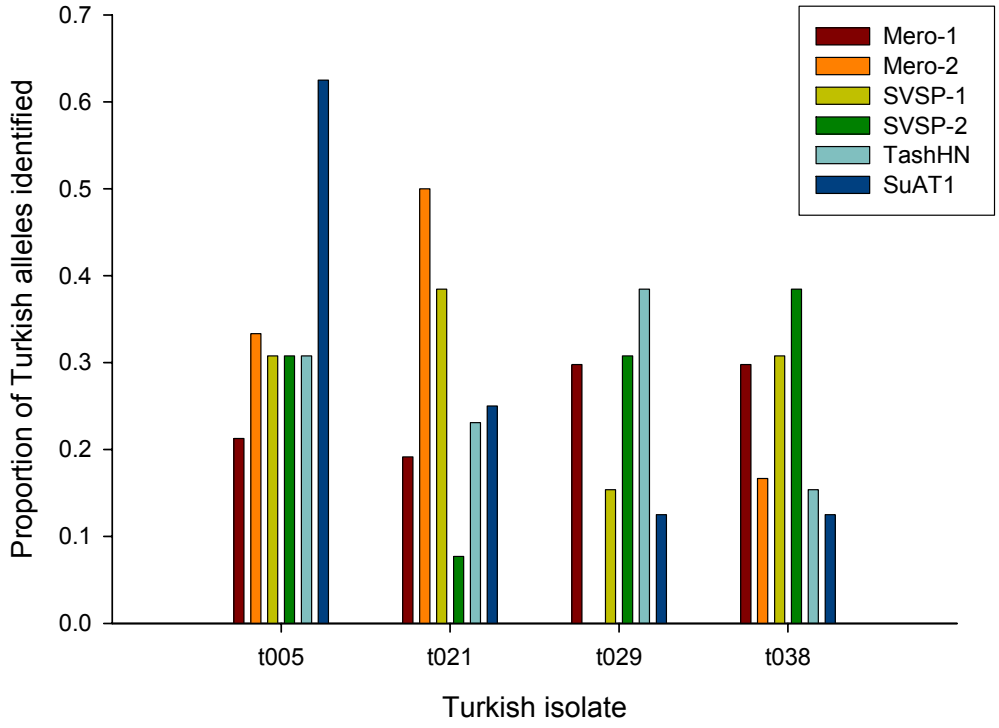
Table 5.5. Summary of unique sequences derived from Tunisia and Turkey

Country	Isolate ID	*mero1*	*mero2*	*SVSP1*	*SVSP2*	*TashHN*	*SuAT$_1$*	mean
Turkey	t005	10	2	4	4	4	5	**3.8**
Number of unique sequences identified within each isolate	t021	9	3	5	1	3	2	**2.8**
	t029	14	-	2	4	5	1	**2.4**
	t038	14	1	4	5	2	1	**2.6**
Subtotal (unique sequences within Turkey)		47	6	13	13	13	8	
Tunisia	w019	1	-	1	-	1	1	**0.67**
Number of unique sequences identified within each isolate	w027	1	-	1	1	1	-	**0.67**
	w030	1	1	1	1	1	-	**0.83**
	w032	1	-	1	1	1	1	**0.83**
	w050	1	1	1	1	1	1	**1.00**
	w062	1	-	1	-	1	1	**0.67**
	w067	1	-	1	1	1	1	**0.83**
	w073	1	1	1	1	1	1	**1.00**
	w101	1	1	1	1	1	1	**1.00**
	w102	1	-	1	-	1	1	**0.67**
Subtotal (unique sequences within Tunisia)		10	4	10	7	4	7	
Total number of unique sequences identified		**57**	**10**	**22**	**20**	**15**	**15**	

Figure 5.3. Proportion of Turkish alleles identified in each isolate

Several alleles were identified at each of the six loci over the four highly heterogeneous Turkish isolates described in Table 5.2. For each locus, the proportion of the total alleles in the population identified in each DNA sample was calculated.

Figure 5.3. Proportion of Turkish alleles identified in each isolate



5.3.2. Measurement of PCR error

In order to determine the error rate of PCR amplification, cloning and sequencing, a preparation of the Ankara C9 (genome clone) of the parasite was included in the panel of template DNA used to amplify the six genes of interest. By comparing the previously published genomic sequence with the results generated using the C9 clone, the frequency of PCR related sequencing errors in this study could be accurately calculated. Actual sequencing errors were effectively ruled out since each nucleotide had greater than '2x' coverage. Five of the six genes of interest were successfully amplified and cloned using this preparation, with the exception of *SuAT1* where insert-containing plasmids were not identified. Two separate clones of C9 were sequenced for the *mero1* (TA13810) gene and a single clone was sequenced for each of the other four amplifying genes. The length of the sequenced amplicon and the number of nucleotide mismatches with respect to the genome sequence for each gene are detailed in Table 5.6. The number of errors varied between zero in the shorter sequences up to a maximum of two in the longer sequences. Previous micro- and mini-satellite genotyping (Chapter Two) identified the Tunisian samples as being clones, containing a single haploid genotype. Therefore, it was assumed that a single allelic sequence would be obtained for each gene from each of these samples. To provide evidence for this assumption, two clones of the *mero1* gene were generated and fully sequenced for seven of the ten Tunisian samples. The quantity of mismatches between pairs of clones could therefore inform on the amount of PCR errors and for this reason these values are also included in Table 5.6. When this dataset was considered as a whole, 15 erroneous nucleotides were identified over a total length of 15,881 bases of DNA sequenced. This indicated that, on average there were 0.94 errors per kilobase, which was equivalent to a single PCR error per 1,064 bases.

In total 57 unique allelic sequences were identified for the *mero1* gene from the Tunisian and Turkish isolates with a consensus length of 654 bp. The number of PCR errors and hence sequencing mistakes occurring over the entire *mero1* allelic dataset was estimated at 35 in total (Table 5.7.). A simple methodology was used to identify and correct for a proportion of this error and a summary of the pertinent calculations are presented in Table 5.7. First, the multiple DNA alignment was reviewed and triplets of nucleotides representing codons were identified. Each codon was analysed in turn, assessing the nucleotide composition of that codon across all sequences. If a point mutation was identified in a single sequence, and all the other alleles were completely conserved as regards this codon then this nucleotide was considered to be a putative PCR error; this process is illustrated in Figure 5.4. Across the 57 *mero1* sequences of 654 nucleotides, a

Table 5.6. Determination of PCR error rate

A series of paired putatively identical sequences were aligned in order to determine the PCR error rate. 15 mismatched nucleotides were identified over a total length of 15,881 bases of PCR amplified sequence indicating that, on average, there were 0.94 errors per kilobase, equivalent to a single PCR error per 1,064 bases.

Table 5.6. Determination of PCR error rate

Gene name	GeneDB ID	Hypothetically identical sequences		Sequence length (bases)	Comparative length (bases)	Number of mismatches	Errors / kb
mero1	TA13810	C9 clone 1	GeneDB	660	660	0	0
mero1	TA13810	C9 clone 2	GeneDB	660	660	0	0
mero2	TA20615	C9 clone	GeneDB	1,466	1,466	2	1.36
SVSP1	TA16025	C9 clone	GeneDB	1,683	1,683	2	1.19
SVSP2	TA17485	C9 clone	GeneDB	1,233	1,233	1	0.81
TashHN	TA20090	C9 clone	GeneDB	999	999	0	0
mero1	TA13810	w019 clone 1	w019 clone 2	657	1314	1	0.76
mero1	TA13810	w027 clone 1	w027 clone 2	657	1314	0	0
mero1	TA13810	w030 clone 1	w030 clone 2	660	1320	3	2.27
mero1	TA13810	w050 clone 1	w050 clone 2	660	1320	2	1.51
mero1	TA13810	w062 clone 1	w062 clone 2	660	1320	1	0.76
mero1	TA13810	w073 clone 1	w073 clone 2	660	1320	3	2.27
mero1	TA13810	w101 clone 1	w101 clone 2	660	1320	0	0
TOTAL				**-**	**15,881**	**15**	**0.94**

Table 5.7. PCR error estimation for *mero1*

The PCR error rate calculated in Table 5.6. was used to estimate the total number of PCR errors in the *mero1* allelic dataset. Putative PCR errors were identified based on the methodology illustrated in Figure 5.4. Using this technique, no putative errors were identified in the C9 clone genome sequence and a conservative number of PCR errors was computed across the *mero1* allelic dataset.

Table 5.7. PCR error estimation for *mero1*

Number of sequences	57	
Length of consensus sequence (bp)	654	
Number of confirmed SNPs (in C9 / GeneDB)	0	
Total estimated PCR errors (@ 0.94 per kb)	35.21	
Observed SNPs in otherwise conserved codons	26	
Estimated PCR errors per allelic sequence	0.618	
Observed SNPs per allelic sequence in otherwise conserved codons	0.456	*(range 0 – 3, SE +/- 0.090)*

Figure 5.4. Example of PCR error correction

Putative PCR errors were identified as single nucleotide polymorphisms which appeared only once over an allelic dataset in an otherwise completely conserved codon. Nucleotides in the allelic sequences that are identical to C9 genome sequence are represented by a dash (-).

(i) *mero1*

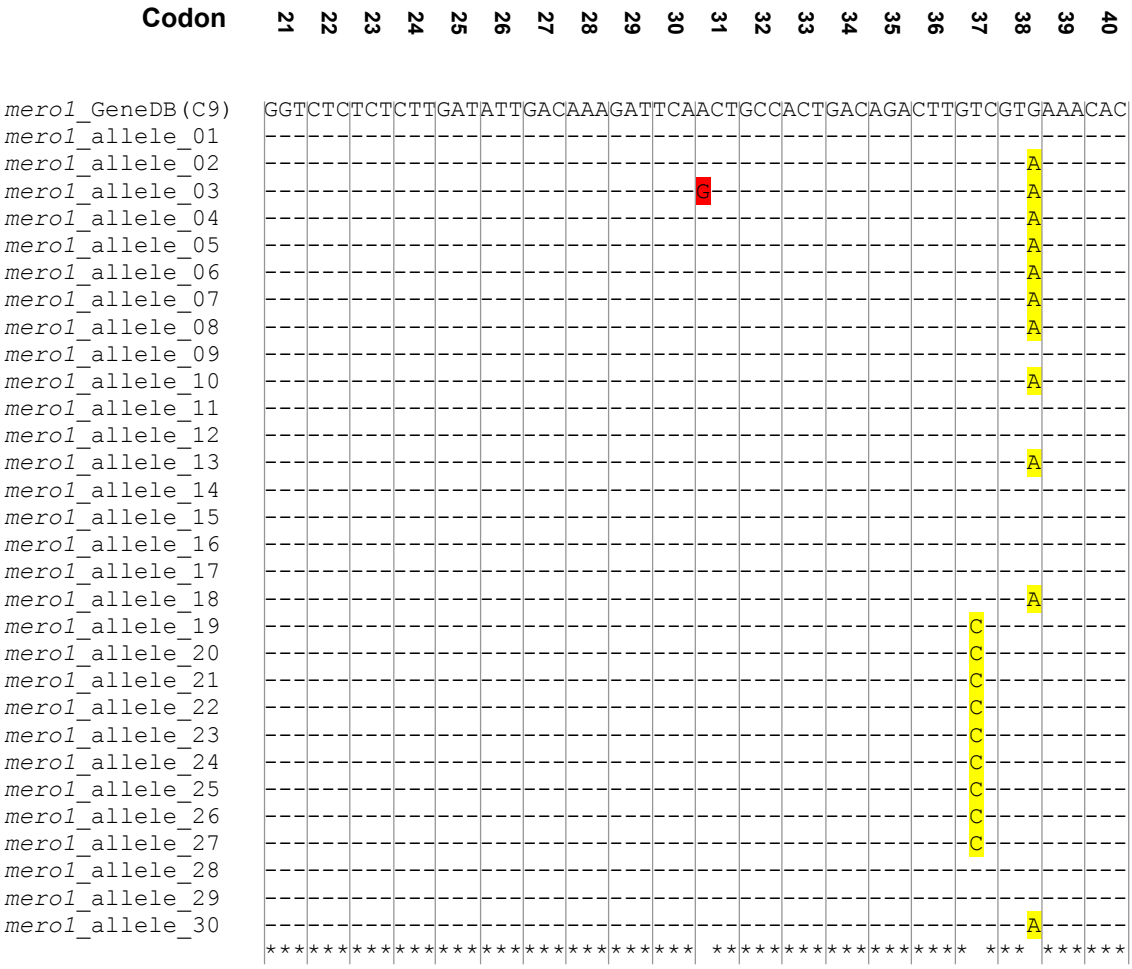
This process was used to correct the large dataset representing alleles of *mero1* (n = 57) and this is illustrated using a subset of sequences. Codons 37 and 38 both include dimorphic sites, where several instances of each nucleotide are represented. Variants from the C9 sequence are highlighted in yellow and this clearly represents genuine sequence diversity. The first nucleotide in codon 31 in allele_03 (G) represents the only polymorphism in that codon over all the alleles and therefore it was identified as a putative PCR error.

(ii) *mero2*

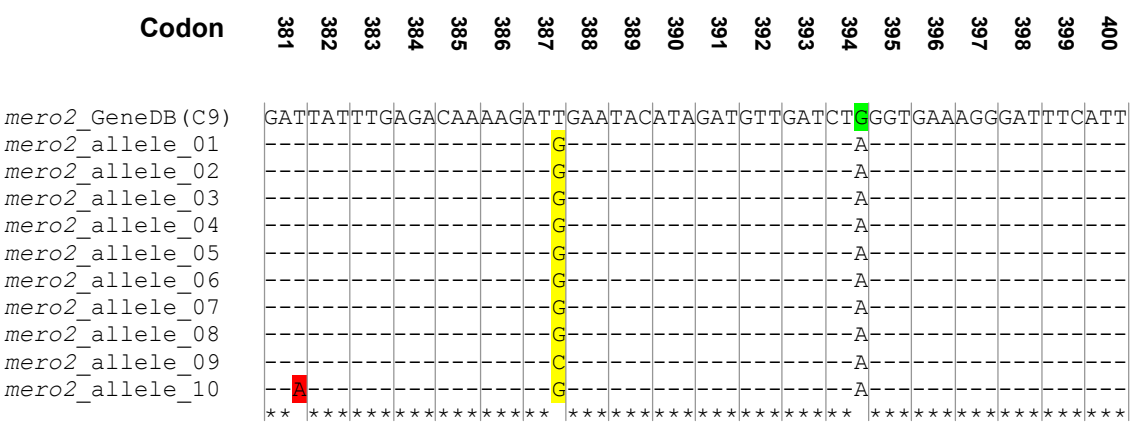
Fewer alleles were sequenced for the other five genes and therefore this methodology was not used. For example, only ten alleles were sequenced for the *mero2* gene. Therefore using this limited dataset, it cannot be determined whether the third site in codon 381 in allele_10 (A) is a genuine polymorphism or represents a PCR error. This is emphasised in codon 394 where the third nucleotide in the C9 sequence is known to be G, however it is invariably A in all the alleles sequenced.

Figure 5.4. Example of PCR error correction

(i) *mero1*



(ii) *mero2*



total of 26 SNPs were identified in otherwise completely conserved codons. The calculated value of 26 was considered conservative but reasonable in comparison to the estimated 35, since a proportion of the PCR errors would have occurred in polymorphic sites ignored by this approach. These polymorphic codons were not included in this process due to the risk removing real SNPs. Both the corrected and uncorrected *mero1* allelic sequence datasets were used in subsequent tests (MK and $d_N d_S$) with virtually identical results observed in each case. For simplicity only the corrected *mero1* results will be presented, except for tests of neutrality (Tajima's and Fu and Li's) where divergent results between the corrected and uncorrected datasets were identified. To support the usefulness of this approach, the C9 allelic sequence of *mero1* from GeneDB (considered to be error-free) was compared to all the other alleles in the dataset using the above protocol and no single nucleotide polymorphisms (corresponding to PCR errors) were identified (Figure 5.4.(i)). Although this is not a definitive test of the approach, encouragingly it did not identify any PCR errors in an allelic sequence known to be error-free (the whole-genome shotgun approach to sequencing did not rely on a PCR amplification step, and sufficient genome coverage was achieved to effectively eliminate sequencing errors). The datasets representing the other genes were not subjected to this error correction method because far fewer numbers of alleles were sequenced. Consequently, estimated PCR error rates bore little relation to the number of suspicious nucleotides identified (data not shown) and it was considered unreasonable to correct the sequences in this manner. Moreover, the technique incorrectly identified putative PCR errors in the genome sequence of other genes, an example of which is illustrated in Figure 5.4.(ii).

5.3.3. Length polymorphism

All but one of the six genes exhibited length polymorphism (Table 5.8.). To identify these non-homologous sites, DNA sequences were translated into protein and aligned with ClustalX. It was determined that the open reading frame was maintained in all cases. For subsequent analysis, DNA sequences were trimmed of sites where gaps were present in the amino acid multiple alignment. The two merozoite candidate genes showed the smallest range of length variation – all alleles of *mero2* were the same size and there were only two size-alleles of the *mero1* insert (660 bp and 657 bp). The smaller of these corresponded to a 3 bp deletion at codon 170 when compared to the genome C9 sequence. The positions of the DNA insertions and deletions for each of the six genes are demonstrated graphically in Figure 5.5. relative to the position of known motifs and domains. Increased length variance was identified in the SVSP genes, with both containing deletions and small insertions, with reference to the C9 allele. In *SVSP1* a large 180 bp deletion was identified

Table 5.8. Summary of sequencing results

For each gene, the length of the coding sequence (CDS) was calculated with reference to the C9 genome sequence. The length of the C9 CDS, which was present in the insert was also calculated. Following multiple sequence alignment, a consensus sequence was generated with the alignment gaps (Figure 5.5.) and the stop codon removed from each sequence. This allowed the percentage sequence coverage of the gene to be calculated with reference to the C9 sequence and represented the proportion of the gene analysed in subsequent tests.

Table 5.8. Summary of sequencing results

Gene name	GeneDB ID	Predicted length of C9 (from GeneDB)			Observed length of C9 insert (bp)	Insert allelic range (bp)	Consensus length *		Coverage of gene †
		CDS (bp)	CDS (codons)	insert (bp)			Nucleotides (bp)	Codons	
mero1	TA13810	690	229	660	660	657 - 660	654	218	95 %
mero2	TA20615	1,473	490	1,466	1,466	1,466	1,461	487	99 %
SVSP1	TA16025	1,683	560	1,683	1,683	1,653 - 1,728	1,494	498	89 %
SVSP2	TA17485	1,233	410	1,233	1,233	1,227 – 1,242	1,194	398	97 %
TashHN	TA20090	999	332	999	999	999 – 1,011	996	332	100 %
SuAT$_1$	TA03135	1,677	558	1,469	-	1,391 – 1,484	1,359	453	81 %

CDS = coding sequence

* in-frame and without stop codon

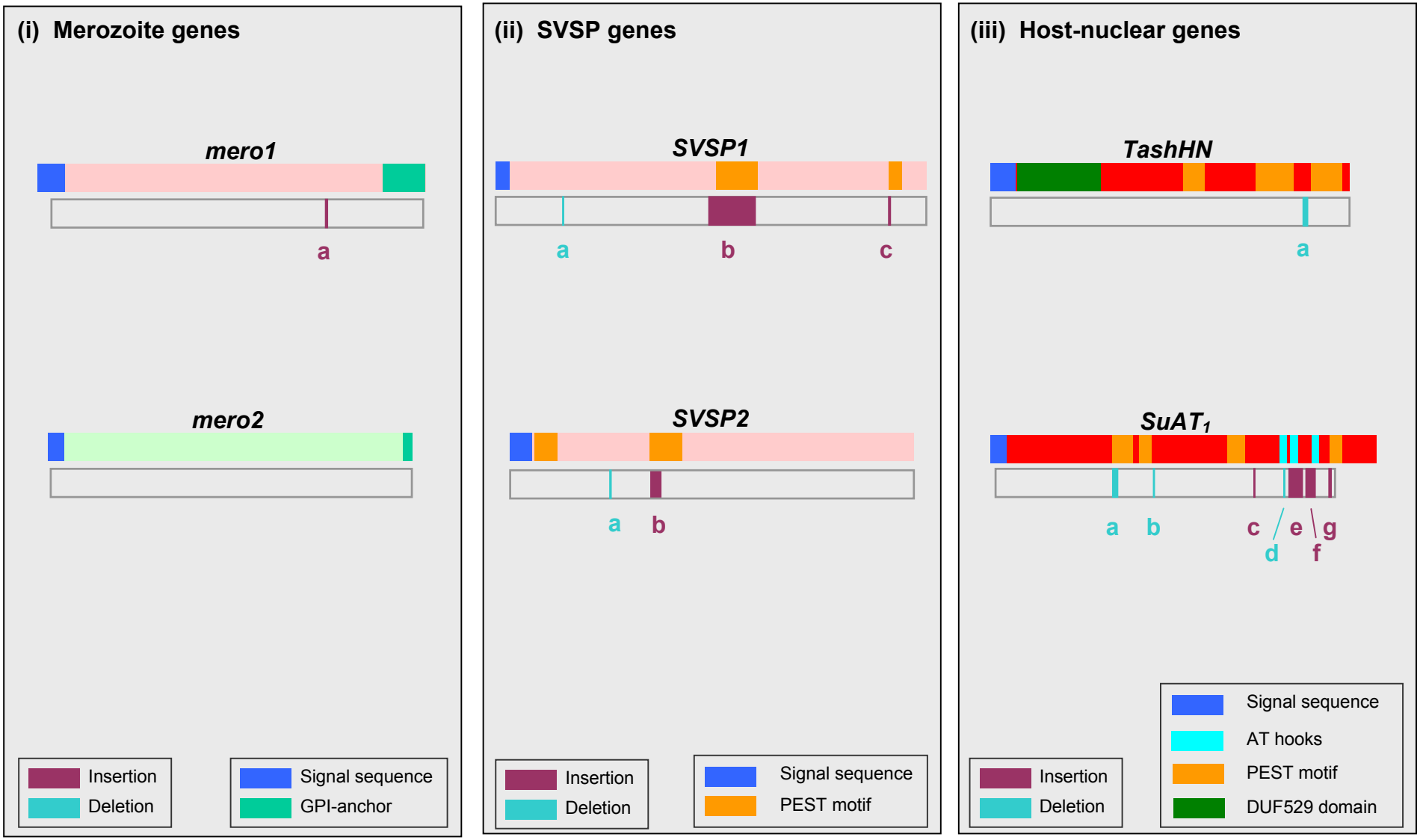
† with reference to the C9 sequence

Figure 5.5. Consensus sequences

In order to analyse allelic sequence data at each locus, sites where insertions or deletions have occurred required identification and removal from the dataset. The coloured graphic for each gene corresponds to the complete coding sequence in the *T. annulata* Ankara C9 genome. Each grey rectangle represents the region of the gene which was PCR amplified and sequenced. Across the alleles, sites of nucleotide insertion relative to the C9 allele are shown in green, while sites of deletion are shown in red.

Gene	GeneDB ID	Locus	Type	Size	Codon
mero1	TA13810	**(a)**	deletion	3 bp	170
mero2	TA20615	**-**	-	-	-
SVSP1	TA16025	**(a)**	insertion	3 bp	between 88 & 89
		(b)	deletion	180 bp	278 – 337
		(c)	deletion	6 bp	512 – 513
SVSP2	TA17485	**(a)**	insertion	3 bp	between 102 & 103
		(b)	deletion	33 bp	144 – 154
TashHN	TA20090	**(a)**	insertion	12 bp	between 290 & 291
SuAT$_1$	TA03135	**(a)**	insertion	21 bp	between 176 & 177
		(b)	insertion	3 bp	between 235 & 236
		(c)	deletion	3 bp	380
		(d)	insertion	3 bp	between 423 & 424
		(e)	deletion	57 bp	430 – 448
		(f)	deletion	39 bp	454 – 466
		(g)	deletion	9 bp	488 – 490

Figure 5.5. Consensus sequences



between codons 278 and 337 corresponding to one of the PEST domains. A smaller 33 bp deletion was identified in the *SVSP2* allele set also within a PEST domain. A 12 bp insertion was identified in *TashHN* corresponding to a four amino acid motif in the C-terminal region of the protein. The largest number of insertions and deletions of any of the genes was identified in the $SuAT_1$ allele set and was focussed in the AT hook region of the encoded protein.

Only homologous nucleotide and amino acid sequence could be used for many of the allelic comparison tests. Therefore, it was important to define the amount of genomic sequence data omitted from the final analyses, and for this reason the percentage of sequence-coverage of each gene was been calculated with respect to the C9 allele (Table 5.8. and Figure 5.5.). For example, in the case of *mero1*, a small portion of the 5' end of the gene was absent due the forward primer being located in the signal sequence and after the gap in the alignment was removed, the residual length of each sequence in the study was 654 bp. This represented 218 codons, 95 % of the 229 codons in the entire C9 allele. Only for *TashHN* was 100 % of sequence analysed with reference to the C9 allele. For the other genes the level varied from 81 % for $SuAT_1$ to 99 % for *SVSP1*. After gap removal, the completely homologous sequences were used to determine polymorphism at the nucleotide level.

5.3.4. Nucleotide polymorphism

To assess the amount of sequence variation present in each sample of alleles, the average number of nucleotide differences (k) and the nucleotide diversity (π) were calculated for each gene and are detailed in Table 5.9. Nucleotide diversity indicates the proportion of nucleotide sites that are different when any two sequences in the set are randomly compared and is equivalent to average heterozygosity for nucleotides across the length of the gene. A sliding window analysis of nucleotide diversity was also undertaken, using a 100 bp frame moving in 25 bp steps and is depicted for each gene in Figure 5.6. This shows a very similar profile of π in both Tunisian and Turkish populations for each gene with the exception of $SuAT_1$. *mero1* shows the highest nucleotide diversity of the six genes in each population and has an overall diversity of 0.0476. Sequence data summarising 99 segregating nucleotides (S) in *mero1* is presented in Figure 5.7. ranging between nucleotide positions 51 and 681. This clearly illustrates the high level of diversity observed between alleles and it is clear that the majority of the sequences have a large number of sites that are different from the C9 genome sequence. When sequences were compared pair-wise, they exhibited on average 31 nucleotide differences ($k = 31.16$,

Table 5.9. Nucleotide polymorphism

Variation among alleles present in each population was assessed by calculating the number of polymorphic sites in the dataset (S), the average number of nucleotide differences (k) and the nucleotide diversity (π) for each gene in each country individually and also across both countries.

Table 5.9. Nucleotide polymorphism

	mero1 †	*mero2*	*SVSP1*	*SVSP2*	*TashHN*	*SuAT$_1$*
Turkey						
Number of sequences	47	6	13	13	13	8
Polymorphic sites (S)	95	87	113	92	18	44
Average number of nucleotide differences (k)	30.63	35.53	41.41	29.01	6.00	13.79
Nucleotide diversity (π)	0.0468	0.0243	0.0277	0.0243	0.0060	0.0101
Tunisia						
Number of sequences	10	4	10	7	4	7
Polymorphic sites (S)	74	60	95	84	12	63
Average number of nucleotide differences (k)	25.93	33.17	30.16	35.38	6.17	28.10
Nucleotide diversity (π)	0.0396	0.0227	0.0202	0.0296	0.0062	0.0207
Total						
Number of sequences	57	10	23	20	17	15
Polymorphic sites (S)	99	107	149	115	20	97
Average number of nucleotide differences (k)	31.16	34.60	42.42	31.08	6.18	41.59
Nucleotide diversity (π)	0.0476	0.0237	0.0284	0.0260	0.0062	0.0306

† calculated using PCR error corrected sequences

Figure 5.6. Nucleotide diversity

For each locus, nucleotide diversity (π) was calculated across allelic sequences, representing Tunisian and Turkish isolates independently. An identical vertical scale was used in all six plots to facilitate comparison between loci. To achieve complete alignment between allelic sequences, which was required for this analysis, gaps in the nucleotide alignment were first removed. This corresponds to the sites of insertion and deletion detailed in Figure 5.5.

Figure 5.6. Nucleotide diversity

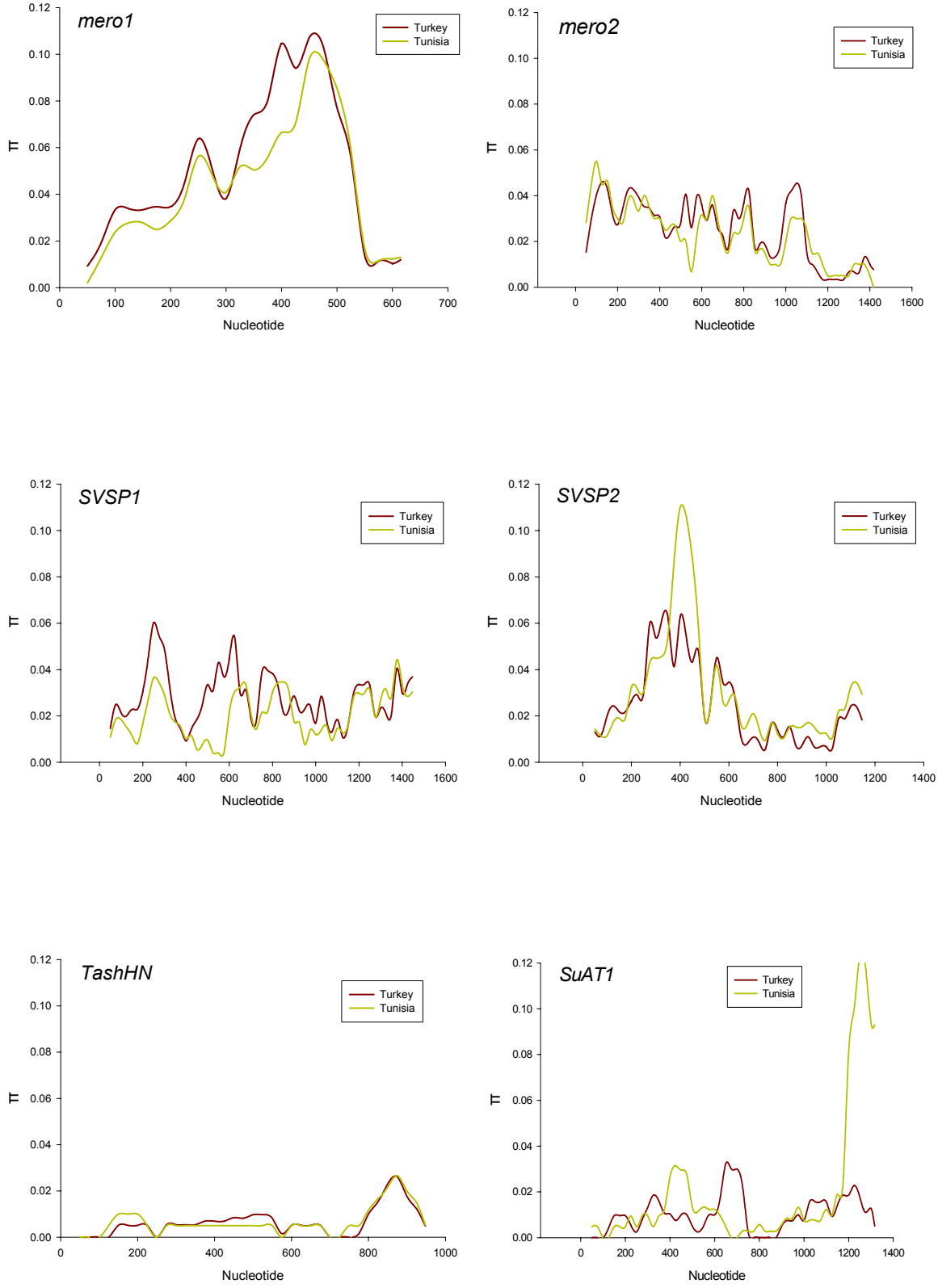
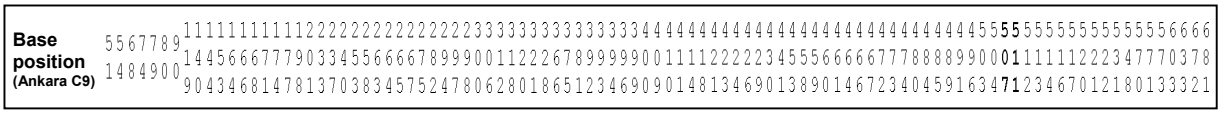


Figure 5.7. Polymorphic nucleotides in *mero1*

99 polymorphic nucleotide sites were identified over the length of the *mero1* gene among alleles from Tunisia and Turkey. A summary of the diversity at each of these sites is presented, using the Ankara C9 sequence (GeneDB) as a reference. Dots in the alignment represent identity with this reference sequence.

Figure 5.7. Polymorphic nucleotides in *mero1*



99 sites across the gene

Table 5.9.). *mero2* had a slightly higher number of average nucleotide differences across the length of the gene, but the value of π was lower (compared to *mero1*) in both Tunisian and Turkish samples at 0.0243 and 0.0227 respectively (Table 5.9.). This apparent discrepancy is explained by the fact that the gene is around twice the length of *mero1*. A contrasting variation in π along the length of the gene can be seen between *mero1* and *mero2* in Figure 5.6. The former gene exhibits a peak in polymorphism between bases 400 and 500, around the position of the 3 bp indel, before diminishing toward the 3' end of the gene and is closely mirrored in each population. *mero2* shows a lower more undulating pattern, with slightly more polymorphism noted in the 5' region. The SVSP genes show an intermediate level of diversity in each population and across the entire sequence set, with overall values for π calculated at 0.0284 for *SVSP1* and 0.0260 for *SVSP2*. Again, a similar distribution along the length of the genes can be seen in each population (Figure 5.6.). *SVSP2* displays maximum diversity towards the 5' end of the gene with very high peak in the Tunisian sample at 400 bp. At this point, the sliding window encompasses a domain of multiple PEST motifs and the flanking sequence of the hypervariable 33 bp deletion. Examination of the multiple sequence alignment confirmed that the diversity in this region was not an artefact generated by block mis-alignment of sequences.

TashHN shows, by far, the least amount of sequence diversity with π being only 0.0062 across the entire sequence set. On average only six nucleotide substitutions will be observed when comparing any two sequences of the 1 kb gene thus the gene is highly conserved in both samples. A small spike in nucleotide diversity can be seen in the region between 800 bp and 900 bp (Figure 5.6.). The 12 bp insertion is located at the centre of this region, which lies between two PEST domains. *SuAT1* generally shows a low level of diversity across the gene within both species, at a level comparable with *TashHN*. However, there is a large increase in diversity in the Tunisian population at around 1,200 bp where the multiple insertion / deletion sites were found. Although π is only 0.0101 and 0.0207 for each individual population, the value rises to 0.0306 when the Tunisian and Turkish populations are combined. This is reflected in the dramatically increased overall value of k, which is 41.59, compared to 13.79 and 28.10 in each country. Low divergence within each sample set and high divergence between sample sets is suggestive of population sub-structuring of this antigen gene with respect to geography. This is highly significant because none of the micro- and mini-satellite markers alone could distinguish isolates from each country. Therefore, at the *SuAT1* locus, the two populations show evidence of divergence with characteristic Tunisian-type and Turkish-type alleles identified. The increase in π around 1,200 bp in the Tunisian population

(Figure 5.6.) is due to the presence of two sequences that are intermediate between Tunisian and Turkish-type alleles. In order to investigate the impact of these differences on population sub-structuring, cluster analysis was performed on the allelic sequences of all the genes under study.

5.3.5. Cluster analysis

Homologous DNA sequences were aligned and clustered using ClustalX allowing a neighbour joining (NJ) tree to be generated for the alleles of each gene, using the method of Saitou and Nei (Saitou and Nei 1987). The main virtue of the NJ method is its efficiency, as it can be used on extensive data sets for which other means of phylogenetic analysis are computationally prohibitive. This method is statistically robust under many models of evolution hence, given sufficient data, the NJ algorithm will reconstruct the true tree with a high degree of probability. The NJ tree for each gene is presented in Figure 5.8. The most extensive NJ tree was generated using the large dataset of *mero1* alleles. Although five of the Tunisian sequences clustered together, the other five were interspersed throughout the dendrogram. Sequences generated from each of the four different Turkish isolates were interspersed with each other for all six genes in the study. That is to say, sequences generated from the same infection did not cluster together indicating an absence of sub-structure within the Turkish population sample. In the case of *mero2* and *SVSP2*, the Tunisian and Turkish isolates are completely interspersed on the dendrograms, without any obvious clustering. In the case of *SVSP1*, there is a significant amount of diversity within the Tunisian isolates and a degree of clustering is observed among five sequences, suggesting an element of sub-structuring. The C9 sequence clusters with the bulk of the Turkish sequences, which are also quite variable and neighbouring two of the Tunisian sequences. A degree of sub-structuring is also evident in the case of *TashHN*. Only one of the ten Tunisian sequences clusters among the bulk of the Turkish sequences, although it is identical to a Turkish sequence isolated from sample t038. A greater degree of variation is evident in the Turkish sample, however because of general limited diversity, the dendrogram is shown at approximately ten times higher resolution. As suggested by the analysis of nucleotide diversity, significant geographical sub-structuring is identified in $SuAT_1$. The eight Tunisian sequences cluster towards the top the dendrogram where two different groups are identified, the larger of which contain six sequences, two of which are identical. In fact, the two Tunisian sequences at the top of the dendrogram are intermediate between the allelic types characteristic each country, and were responsible for the large increase in π at around 1,200 bp in the Tunisian sample set

Figure 5.8. DNA neighbour-joining trees

Neighbour-joining trees were generated using allelic nucleotide sequence corresponding to the six loci of interest. Entries were colour-coded in order to differentiate alleles representing the Tunisian clones and each of the four highly heterogeneous Turkish isolates. Additionally, the *T. annulata* Ankara C9 allelic sequence (from www.genedb.org) is included for each gene.

Figure 5.8. DNA neighbour-joining trees

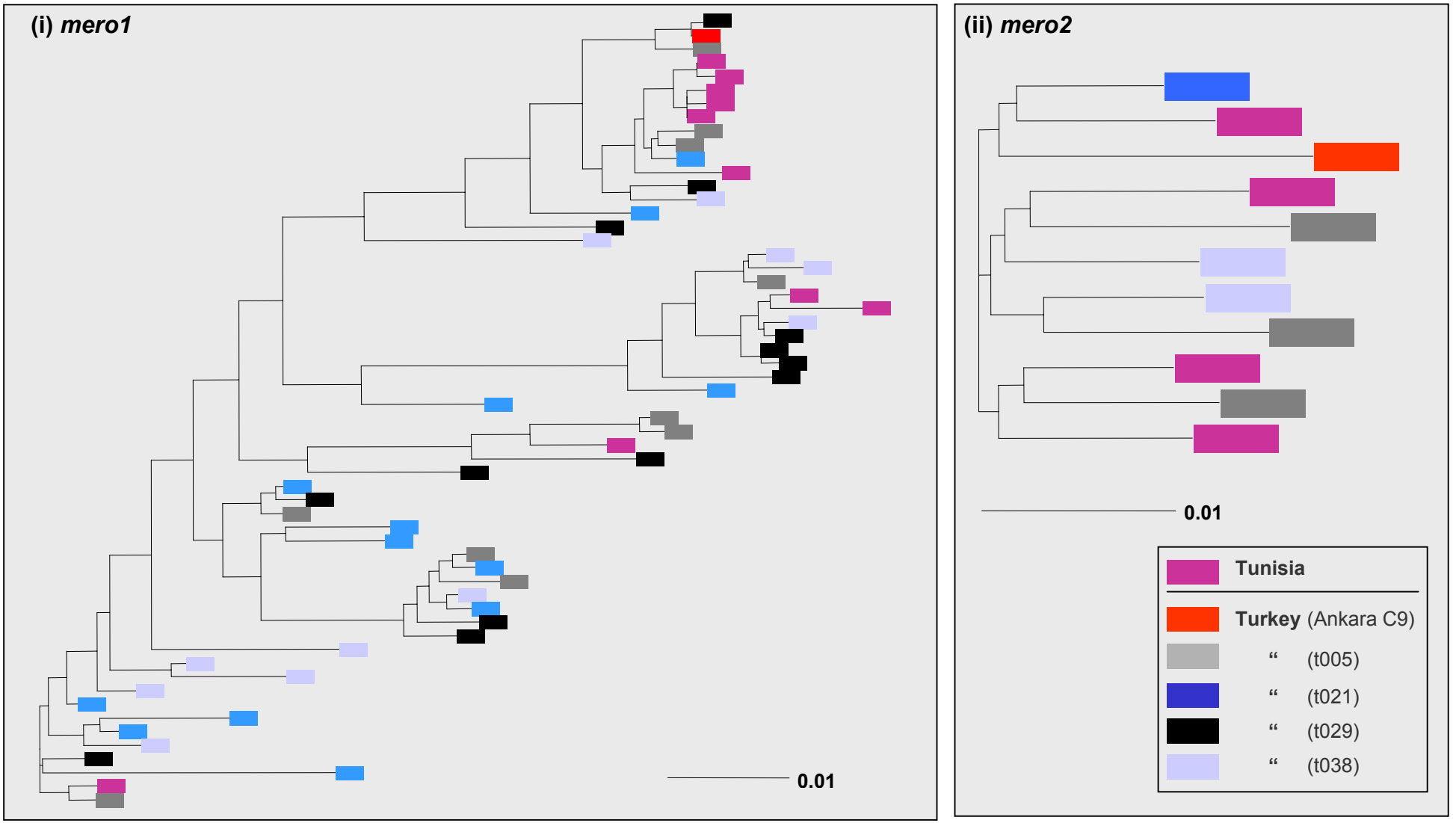


Figure 5.8. DNA neighbour-joining trees (continued)

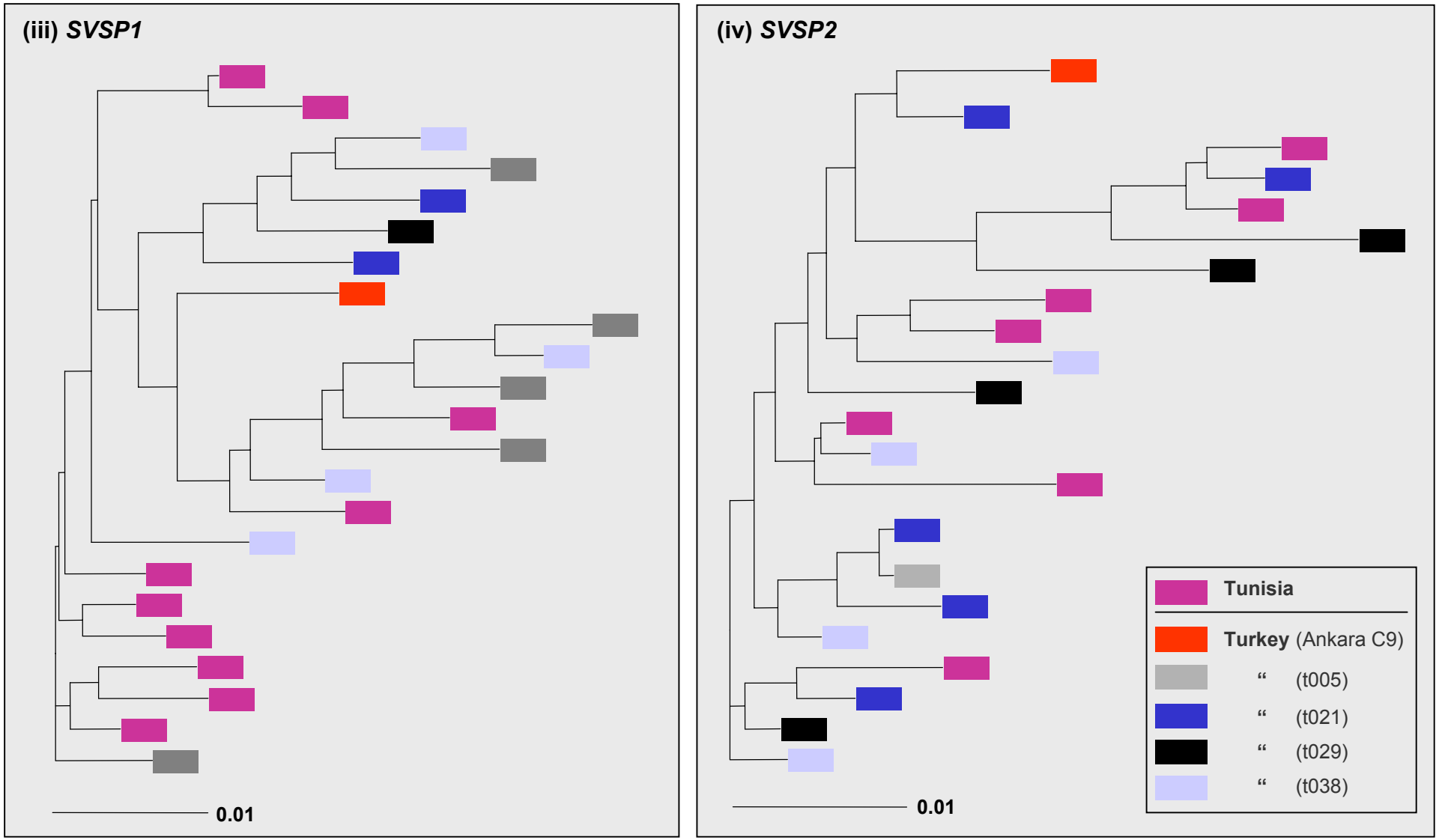
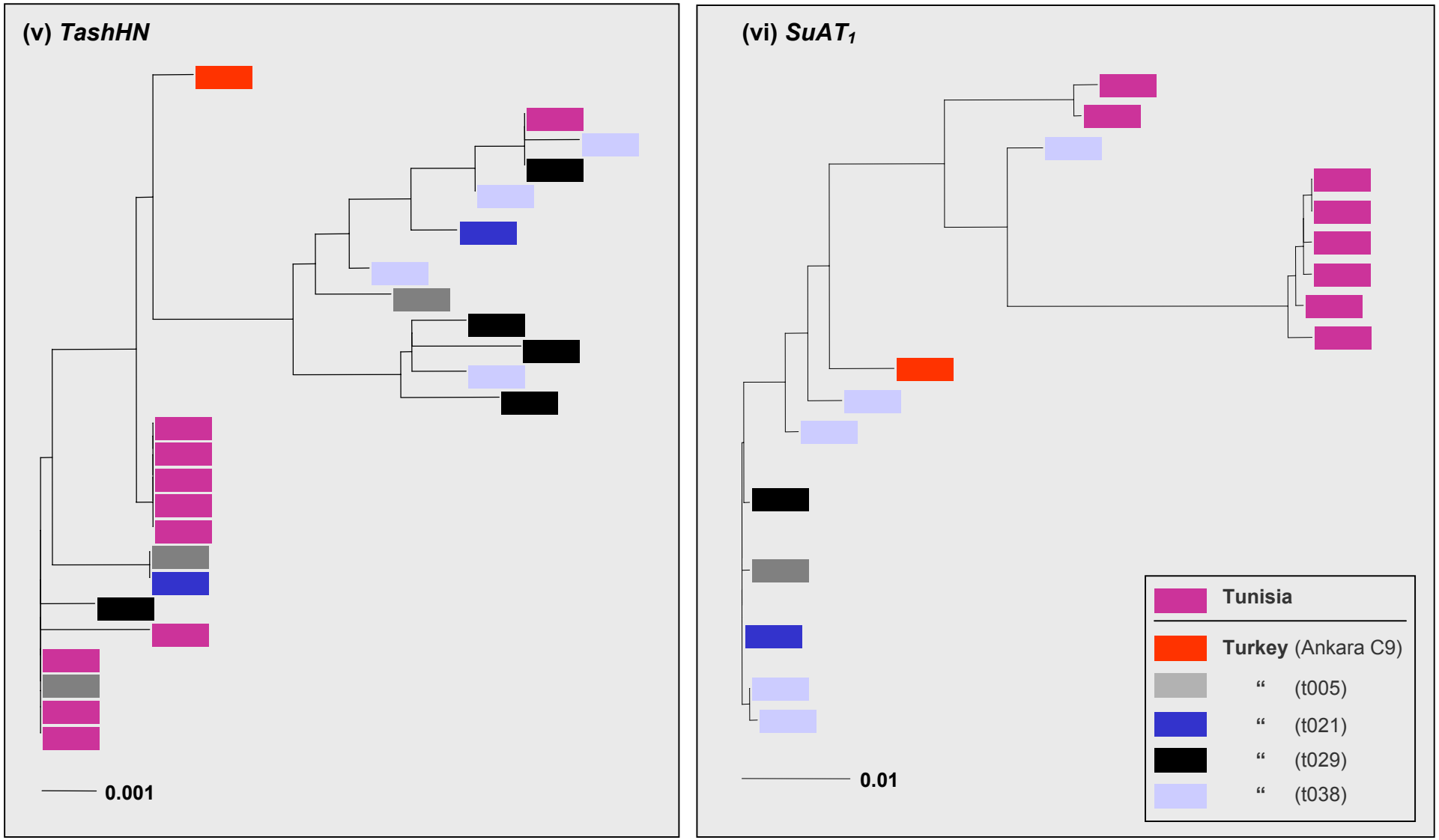


Figure 5.8. DNA neighbour-joining trees (continued)



(Figure 5.6.). A Turkish sequence from animal t005 and the sequence of Ankara C9 (GeneDB) are also intermediate between the allelic types characteristic of each country.

Although nucleotide diversity and cluster analysis provide a useful description of the variation within each gene, it was necessary to characterise the nature of nucleotide substitutions to investigate the type of selective pressure directing the evolution of each of these genes.

5.3.6. Evidence of selection

5.3.6.1. The McDonald-Kreitman test

The McDonald-Kreitman test was used to quantify the pattern of non-synonymous and synonymous polymorphism of *T. annulata* alleles compared to the orthologous gene within a closely related species. The results of these analyses are presented in Table 5.10. The *T. parva* sequence was available for all six loci, with additional sequences of *mero1* available for *T. buffeli*, *T. sergenti* (Ikeda stock) and *T. sergenti* (Chitose stock). An analysis of the published *TaMS1* allelic sequences (Katzer *et al.* 1998; Gubbels *et al.* 2000b) was also undertaken principally to compare with the results of the two merozoite candidate genes. The *TaMS1* data represents a disparate collection of sequences from many parts of the world and therefore the results must be interpreted with some caution.

The merozoite candidate genes *mero1* and *mero2* showed a contrasting pattern of variation in comparison with *T. parva*. *mero1* exhibited an excess of non-synonymous substitutions – 55 non-synonymous vs 51 synonymous substitutions within *T. annulata* compared to 24 non-synonymous vs 42 synonymous substitutions between the species (Table 5.10.). The skewedness of the ratios was reflected by a neutrality index of 1.887 with an accompanying *p* value of 0.059 generated by Fisher's exact test. In other words, there was less than a 6 % probability that the excess of non-synonymous in *T. annulata* was caused by chance. When *mero1* was compared to the *T. buffeli* and *T. sergenti* sequences, non-synonymous substitutions were similar to synonymous substitutions and a neutrality index slightly greater than one was achieved in each case. However, none of these comparisons were statistically significant with high *p* values indicating random effects may explain the results. *mero1* showed a similar pattern to *TaMS1*, where there was an excess of non-synonymous substitutions observed within *T. annulata*. Interestingly, although the proportion of non-synonymous to synonymous substitutions was higher in *TaMS1*, the neutrality index was lower and the *p* value of Fisher's test indicated less significance. This was attributed to an increased proportion of fixed non-synonymous to synonymous

Table 5.10. McDonald-Kreitman test

The McDonald-Kreitman test was used to compare the pattern of non-synonymous (Nsyn) and synonymous (Syn) substitutions within *T. annulata* alleles against an orthologous gene in another *Theileria* species. A neutrality index of greater than one indicated an excess of intra-specific non-synonymous substitutions, while a value of less than one indicated a deficit; values close to one indicated no difference. Fisher's exact test of significance was used to test the null hypothesis that there was no difference in the proportion of intra-species Nsyn / Syn substitutions to the proportion of inter-species Nsyn / Syn substitutions.

Table 5.10. McDonald-Kreitman test

Gene name	GeneDB ID	Species *	Nucleotides	Number of *T. annulata* sequences	Polymorphic changes within *T. annulata*		Fixed differences between species		Neutrality index	*p* value (Fisher's exact test)
					Syn	Nsyn	Syn	Nsyn		
TaMS1	*TA17050*	*T. parva*	*741*	*119*	76	106	52	55	1.319	0.272
mero1 †	TA13810	*T. parva*	654	58	51	55	42	24	1.887	0.059
mero1 †	TA13810	*T. sergenti (Ikeda)*	654	58	50	55	99	104	1.047	0.904
mero1 †	TA13810	*T. sergenti (Chitose)*	654	58	50	55	103	99	1.144	0.631
mero1 †	TA13810	*T. buffeli*	654	58	50	55	107	103	1.143	0.633
mero2	TA20615	*T. parva*	1,461	11	72	34	117	149	0.371	0.000033
SVSP1	TA16025	*T. parva*	1,494	23	42	108	113	361	0.805	0.329
SVSP2	TA17485	*T. parva*	1,194	21	36	69	127	292	0.834	0.480
TashHN	TA20090	*T. parva*	996	16	11	9	84	139	0.494	0.153
SuAT$_1$	TA03135	*T. parva*	1,359	16	16	49	174	353	1.510	0.205

* species with which *T. annulata* sequences compared, Syn = number of synonymous changes, Nsyn = number of non-synonymous changes

† calculated using PCR error corrected sequences

substitutions between the species. In contrast to *TaMS1* and *mero1*, *mero2* demonstrated an excess of synonymous substitutions within *T. annulata* with an inverted ratio observed between species. This resulted in a low neutrality index of 0.371 indicating purifying selection was operating on this gene within *T. annulata*. The extreme nature of the amino acid sequence conservation led to a very low *p* value for Fisher's exact test (0.000033), which was highly significant.

Both SVSP genes displayed a similar profile when the McDonald-Kreitman test was applied. The ratio of non-synonymous to synonymous substitutions within *T. annulata* and between *T. parva* was 3:1 for *SVSP1* and 2:1 for *SVSP2* for each comparison. There was no significant difference in this ratio observed between the species compared to within *T. annulata*, consistent with a genuinely neutral pattern of evolution both within and between species. The neutrality indices for each of these genes were 0.805 and 0.834, suggesting that the sequences were operating close to neutrality, with a slight tendency towards purifying selection. It must be borne in mind that hypervariable regions in both of these genes were omitted from this comparison. However, these represented only around 10 % of the *SVSP1* sequence and less than 5 % of *SVSP2* as was demonstrated in Figure 5.5.

Similar to the merozoite candidates, the parasite-encoded host nuclear genes showed two contrasting patterns of variation. Both genes did, however, display a limited amount of diversity within *T. annulata* but a large amount of interspecies divergence, distinguishing them from the four other genes under study. In the case of *TashHN*, there were less non-synonymous than synonymous substitutions in the *T. annulata* population (9 vs 11), whereas between species there was an excess of non-synonymous differences. This indicated that the gene had been conserved within *T. annulata* whilst divergence between the species has occurred. The results approached statistical significance with an associated *p* value of 0.153. In contrast, $SuAT_1$ showed a high proportion of non-synonymous substitutions at both an intra- and inter-species level. The proportion within *T. annulata* itself was higher, however, indicating that diversifying selection was operating on the gene and this is mirrored in the neutrality index of 1.510. Three factors must be considered, however, when assessing the significance of this particular result – (1) there was strong evidence of population sub-structuring between Tunisian and Turkish $SuAT_1$ sequences of *T. annulata*, (2) the sample size in each population was relatively small (Turkey n = 8 (& C9), Tunisia n = 7) and (3) only around 80 % of the entire coding sequence was included in the analysis, due to non-homologous tracts in the hypervariable region.

To summarise, the results of the McDonald-Kreitman test suggested that similar to *TaMS1*, *mero1* is under diversifying selection within *T. annulata*, in stark contrast to *mero2*, which was shown to be highly conserved and under purifying selection. Equivocal results were generated for the two SVSP genes, with a large amount of both synonymous and non-synonymous substitutions observed within *T. annulata* and between *T. parva* and *T. annulata*. Additionally, the results indicated that *TashHN* is relatively conserved within *T. annulata* and divergent between species whereas *SuAT1* appears to be both divergent between species and divergent in *T. annulata*. To gain further insight into the relative influence of diversifying and purifying selection on these genes, $d_N d_S$ analysis was undertaken.

5.3.6.2. $d_N d_S$

$d_N d_S$ analysis provides a means for characterising the influence of nucleotide polymorphism on amino acid diversity across a set of homologous sequences. The nucleotide composition of each codon may be analysed across a set of alleles to determine whether DNA diversity results in amino acid conservation or polymorphism. If a gene is evolving neutrally, then an equivalent number of non-synonymous and synonymous changes are expected ($d_N = d_S$). Codons that exhibit $d_N > d_S$ are more prone to amino acid altering substitutions and are identified as being under the influence of positive selection. In contrast, codons that exhibit $d_N < d_S$ tend to encode invariant amino acids and are therefore considered to be under the influence of negative or purifying selection. A mean $d_N d_S$ value may be calculated over the length of each gene to reflect the relative abundance of positively and negatively selected codons. As described in Section 5.2.4.3., the software package HyPhy was used to determine the global ratio of d_N to d_S and also to analyse each gene on a codon-by-codon basis. The results for all six genes in the present study are presented in Table 5.11., together with the *T. annulata* / *T. parva* interspecies values calculated using PAML (Section 4.2.1.). The 87 published *TaMS1* sequences were also analysed to give a $d_N d_S$ ratio of 0.5759, which was greater than that of any of the other genes analysed. The greatest number of statistically significant positively selected sites, was displayed by this gene, with ten sites identified where $p < 0.1$, compared to 26 negatively selected sites. To determine the distribution of positively ($d_N > d_S$) and negatively ($d_N < d_S$) selected sites along the length of each gene, $d_N - d_S$ values calculated on a codon-by-codon basis were plotted for each gene (Figure 5.9.). For *mero1*, a cluster of positively selected sites was observed around codon 150 (Figure 5.9.(i)), coinciding with the region of high nucleotide diversity (Figure 5.6.) at around nucleotide 450. This region is adjacent to the 3 bp indel identified across *mero1* alleles. Negatively selected

Table 5.11. $d_N d_S$ analysis

The allelic dataset for each gene was analysed using HyPhy (Section 5.2.4.3.). Using a maximum likelihood method, the number of codons where $d_N > d_S$ (positively selected sites) and the number of codons where $d_N < d_S$ (negatively selected sites) was calculated using two different levels of significance ($p < 0.25$ and $p < 0.1$). The ratio of d_N to d_S was also calculated across the length of each gene and this is indicated alongside the *T. annulata* / *T. parva* interspecies value calculated using PAML (Section 4.2.1.).

Table 5.11. $d_N d_S$ analysis

Gene name	GeneDB ID	Number of sequences analysed	Number of codons analysed	Positive sites (*p < 0.25*)	Negative sites (*p < 0.25*)	Positive sites (*p < 0.1*)	Negative sites (*p < 0.1*)	*T.a.* allelic $d_N d_S$ (HyPhy)	*T.a.* vs *T.p.* $d_N d_S$ (PAML)
TaMS1	*TA17050*	*87*	*274*	*17*	*44*	*10*	*26*	*0.5759*	*0.2751*
mero1 †	TA13810	58	218	5	37	3	28	0.4131	0.1638
mero2	TA20615	11	487	1	58	0	32	0.1331	0.2560
SVSP1	TA16025	23	498	8	27	2	16	0.5466	0.4037
SVSP2	TA17485	21	398	5	23	2	14	0.4643	0.3118
TashHN	TA20090	16	332	0	5	0	3	0.2111	0.2646
SuAT$_1$	TA03135	16	453	0	9	0	4	0.5024	0.2192

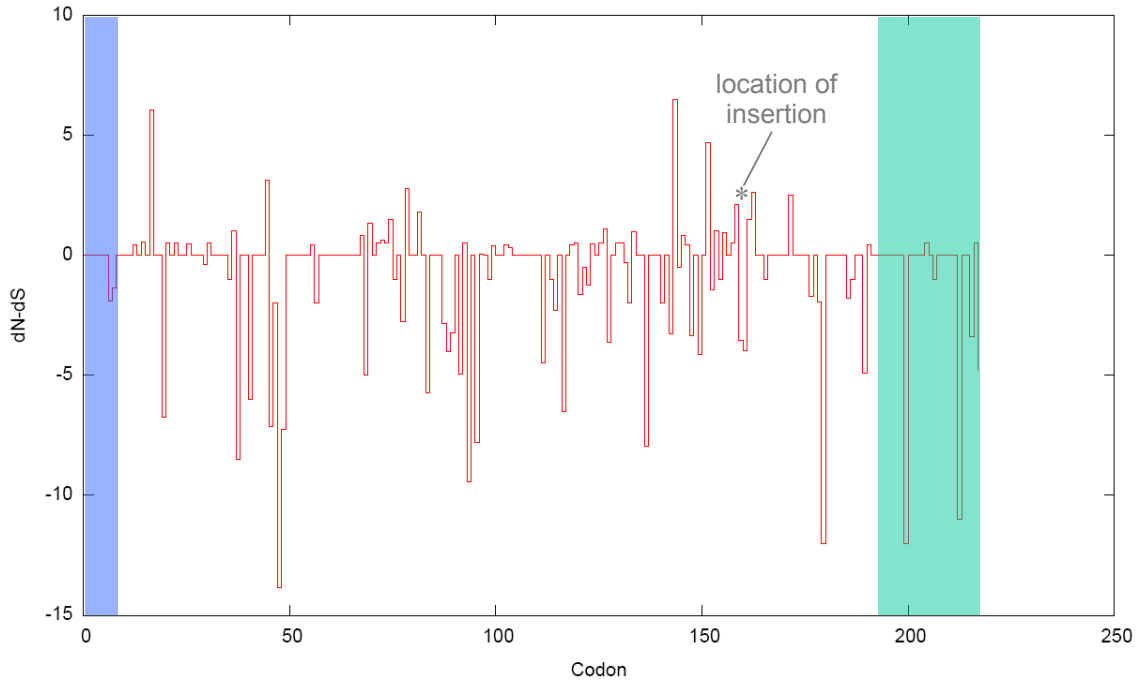
† calculated using PCR error corrected sequences

Figure 5.9. $d_N d_S$ plots

For each locus, allelic sequences were compared to determine the non-synonymous to synonymous substitution rate ($d_N d_S$) across the length of the gene, analysing each codon independently. Values of $d_N - d_S$ above zero indicated positive selection, whereas values beneath zero indicated purifying selection. For each locus, a graphic illustrates the location of various peptide motifs in the translated gene product. A different vertical scale is used in each diagram, which was determined in each case by the minimum and maximum value of $d_N - d_S$ over the length of the sequence.

Figure 5.9. $d_N d_S$ plots

(i) ***mero1***



(ii) ***mero2***

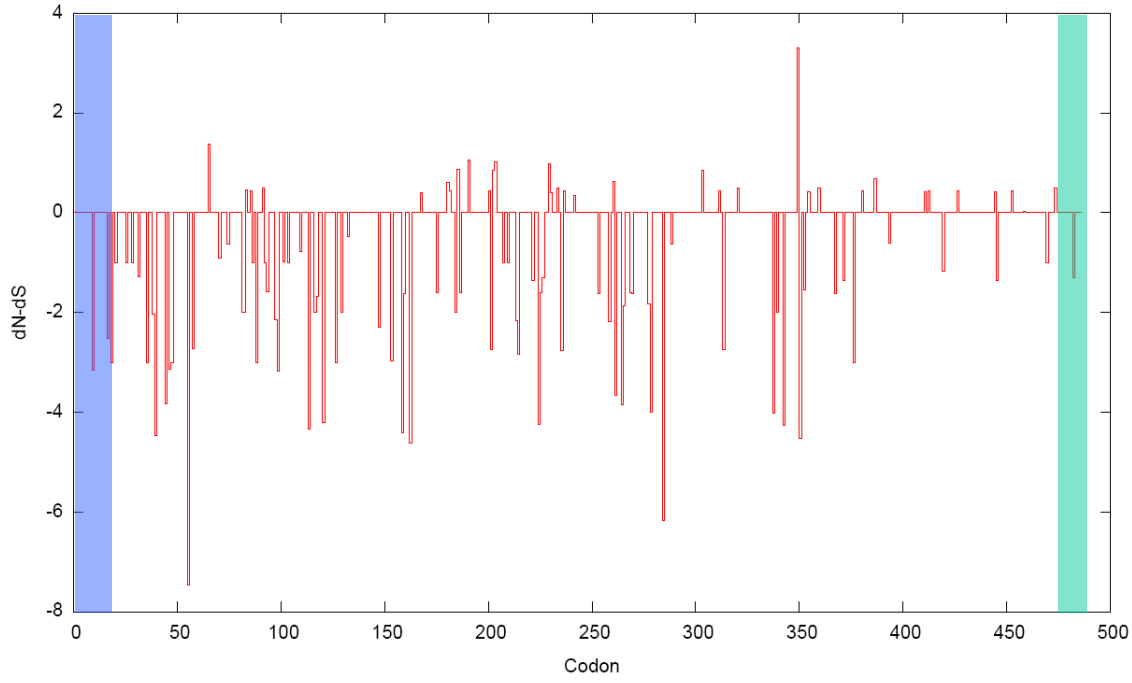
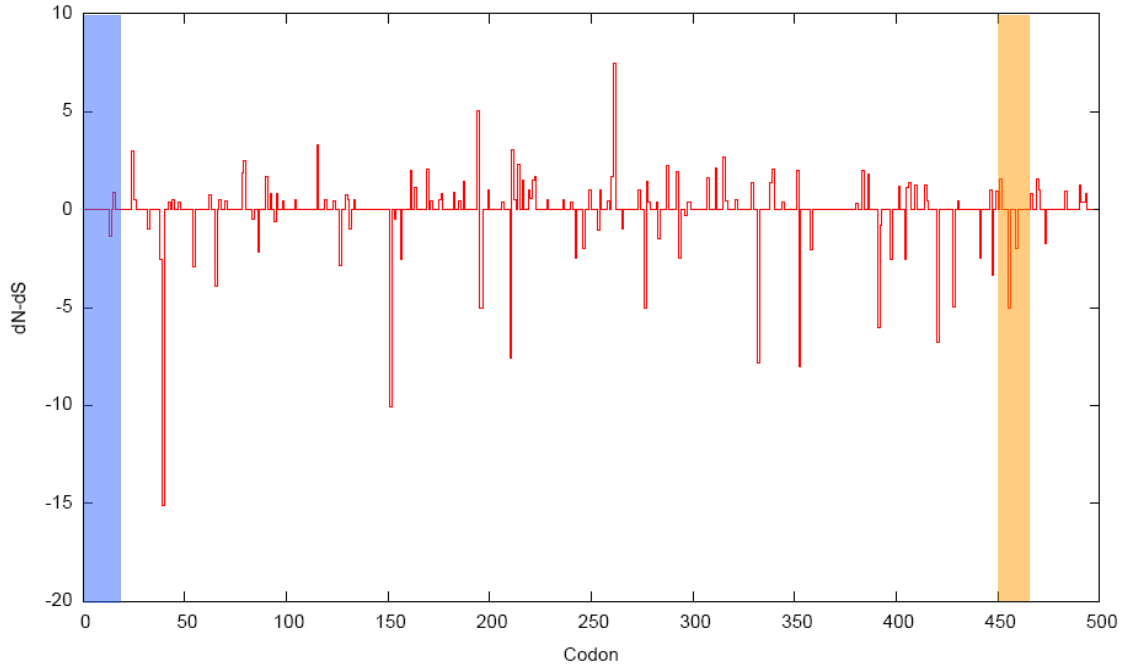


Figure 5.9. $d_N d_S$ plots (continued)

(iii) SVSP1



(iv) SVSP2

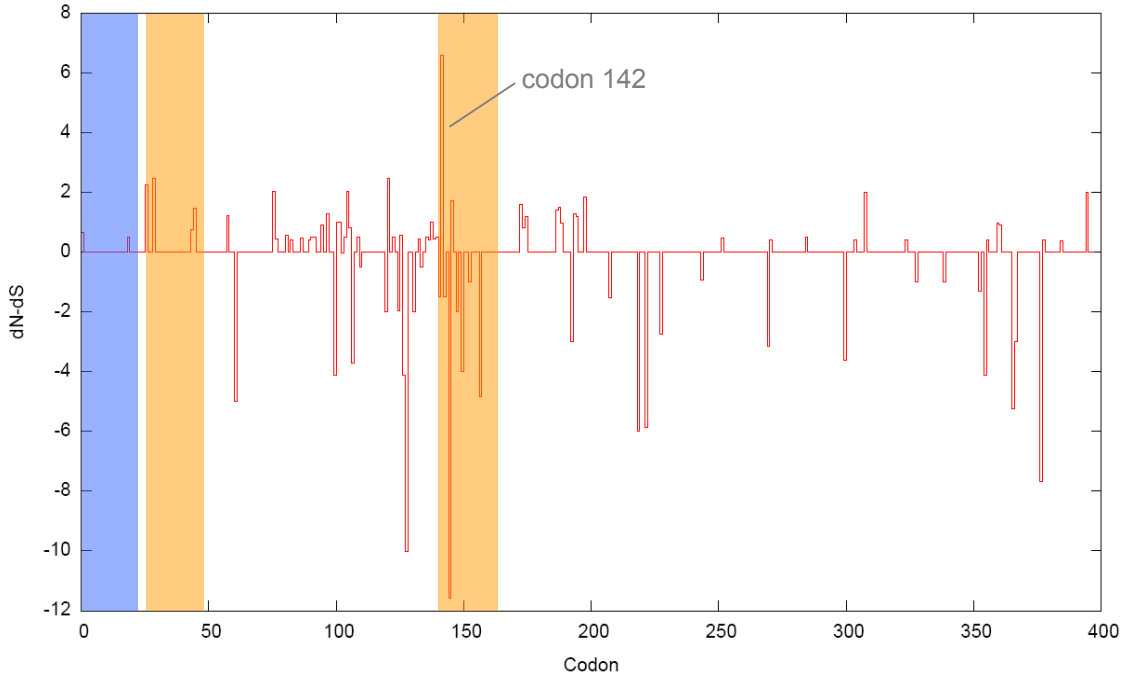
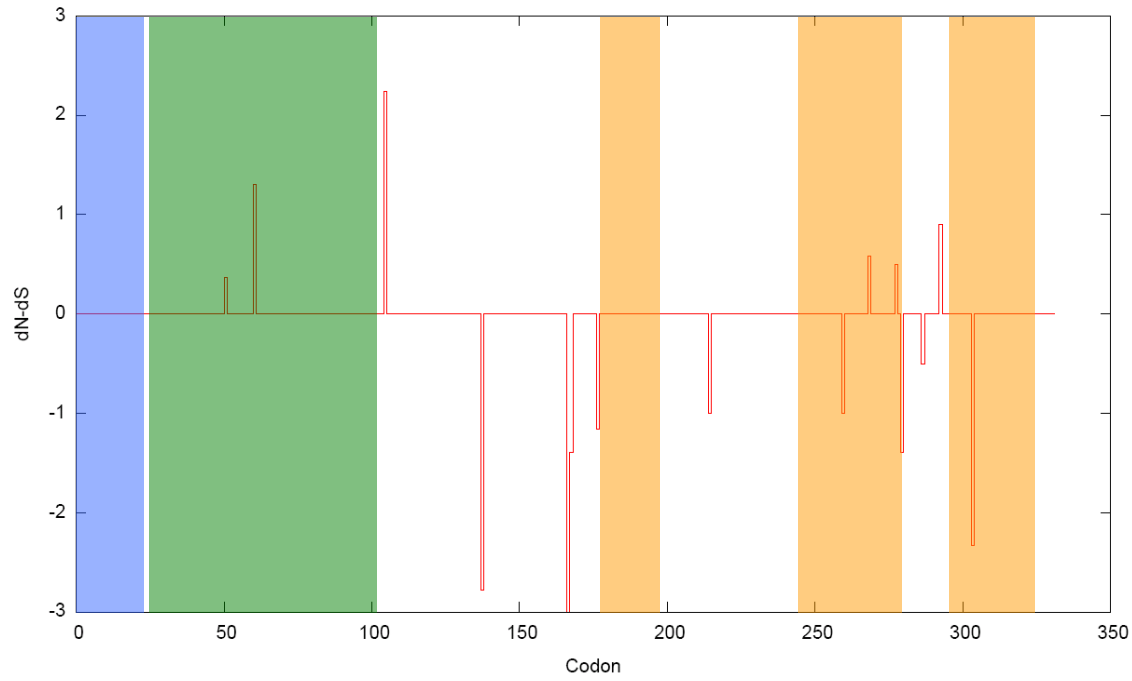
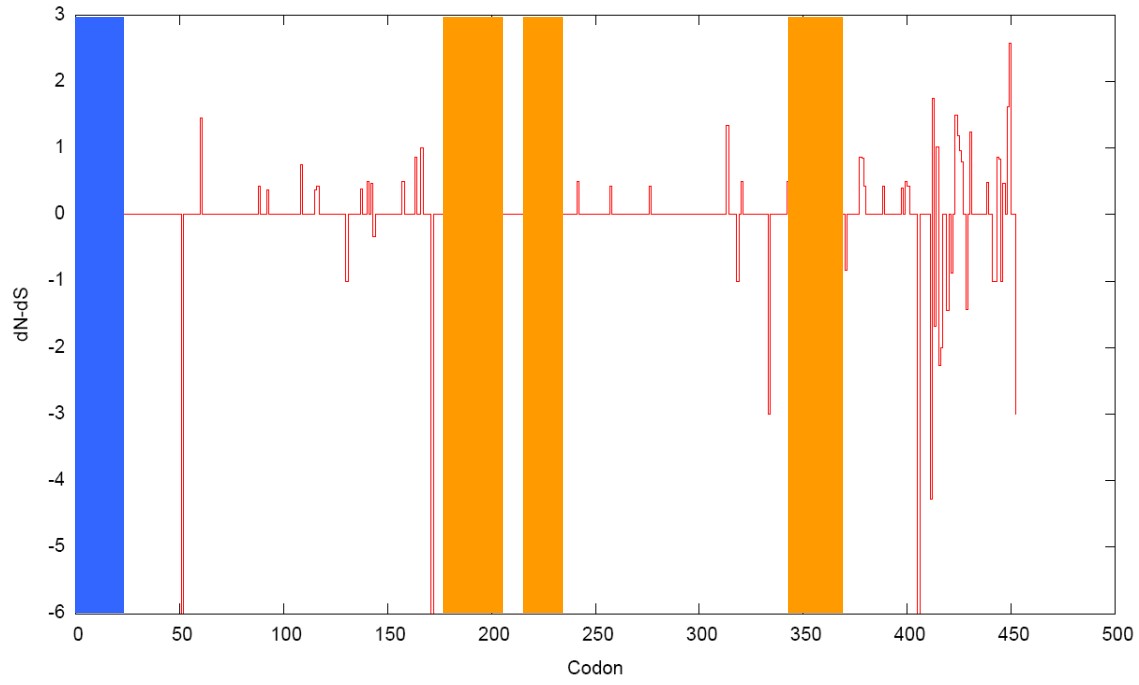


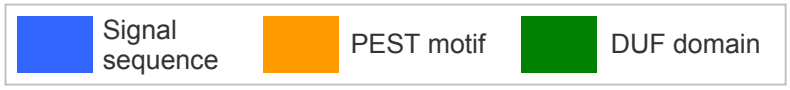
Figure 5.9. $d_N d_S$ plots (continued)

(v) *TashHN*



(vi) *SuAT$_1$*





sites, i.e. those under purifying selection, included the 5' and 3' extremes where the signal and GPI motifs are encoded. Interspersed between the sites of strong positive and negative selection, low positive and low negative peaks indicated neutral codons. Absence of a peak at a codon represents complete conservation of all three encoding nucleotides across the set of alleles. In parallel with the results of the McDonald-Kreitman test, the merozoite candidate genes displayed contrasting results. A high intra-species ratio of 0.4131 was displayed for *mero1*, while a much lower value of 0.1331 was calculated for *mero2*. When compared with the inter-species values, the value for *mero1* has more than doubled while the value for *mero2* has almost halved. This indicated that within *T. annulata*, *mero1* is under the influence of diversifying selection whereas *mero2* is highly conserved. In common with all the genes in the study, statistically significant synonymous changes are more abundant than non-synonymous changes (Table 5.11.). The signal peptide sequence and the GPI-anchor domain were also strongly conserved in *mero2*, along with a clear general trend toward silent substitutions throughout this gene (Figure 5.9.(ii)). An abundance of negatively selected sites can clearly be seen over the length of the gene with a higher proportion of neutral (or totally conserved) sequence towards the 3' end. This was reflected in the relative numbers of negatively and positively selected sites presented in Table 5.11. At $p < 0.25$, 58 negatively selected codons were identified compared to only one positively selected codon. This clearly indicated that the gene was subject to high level of purifying selection.

Both SVSP genes showed an elevated level of $d_N d_S$ compared to their high inter-species values (Ankara C9 vs *T. parva*). *SVSP1* displayed an abundance of non-synonymous changes along the length of the gene (Figure 5.9.(iii)) and when a *p* value of less than 0.25 was considered, *SVSP1* displayed eight positively selected sites, second only to *TaMS1* in this respect (Table 5.11.). Non-synonymous changes were located towards the 5' end of *SVSP2* (Figure 5.9.(iv)), within the cluster of polymorphism identified in the nucleotide diversity study between 200 and 500 bp (Figure 5.6.). Synonymous substitutions were scattered throughout the molecule, although they were somewhat concentrated in the polymorphic region. Again, the signal peptides in both SVSP genes were conserved, with nucleotide sequences being largely invariant.

As indicated by π, *TashHN* showed a very limited amount of nucleotide polymorphism (Table 5.9.). However, several statistically significant negatively selected sites were identified, whereas positively selected sites were absent (Table 5.11. and Figure 5.9.(v)). The lack of polymorphism could be explained in two ways- (1) very recent evolution of the

locus and / or (2) extreme conservation of the gene where both non-synonymous and synonymous substitutions were selected against. Although slightly counter intuitive, synonymous substitutions may be selected against in some circumstances. For example, if a gene is encoded by a subset of 'optimal' codons associated with high expression levels, when a synonymous substitution occurs, although the amino acid change is silent it is now encoded by a 'non-optimal' codon. This may have a detrimental effect on the level or rate of expression, which may in turn be related to the relative concentrations of tRNA species available (Andersson and Kurland 1990; Dong *et al.* 1996).

A reduced $d_N d_S$ of 0.2111 in *TashHN* relative to the inter-species value supported the results of the McDonald-Kreitman test, where a limited amount of intra-species diversity was observed following comparison with a single sequence of *T. parva*. *SuAT1* displayed slightly unusual, but not altogether unexpected results using $d_N d_S$ (Table 5.11.). Similar to *TashHN*, no codons were identified that were statistically significantly under the effect of positive selection, whereas several negatively selected sites were observed. However, the $d_N d_S$ value of 0.5024 was more than double the inter-species value. When the distribution of substitutions was assessed over the length of the gene, Figure 5.9.(vi), positively selected sites were seen uniformly scattered with a degree of concentration in the 3' region. A concentration of negatively selected sites was also seen at this point, which coincided with the flanking sequence of the hypervariable domain, which was removed from the multiple alignment. Amino acid alignments in the 3' region of the consensus sequence were earlier shown to correspond to different allelic types of *SuAT1* – i.e. Tunisian-type, Turkish-type and an intermediate form. Consequently, localised misalignments are likely to account for a proportion of the positively selected sites in this region and partially explain the high $d_N d_S$ value for this gene.

To summarise, a large number of positively selected codons were identified in *TaMS1*. Similarly, *mero1* contained a cluster of positively selected codons and together with high $d_N d_S$ over the entire sequences, both loci were indicated as being under the influence of positive selection. This was in marked contrast to *mero2*, which was highly conserved at a protein level and therefore under the influence of negative selection. SVSP genes exhibited high $d_N d_S$, associated with an abundance of both synonymous and non-synonymous mutation over the length of the genes. *TashHN* showed very limited polymorphism and was therefore highly conserved at both the DNA and protein level while the analysis of *SuAT1* indicated a high global $d_N d_S$ value but no statistically robust evidence was obtained for selection on any individual codon.

5.3.7. Tests of neutrality

To investigate how variation in the allelic sequences of the six genes may depart from neutrality, Tajima's *D* and Fu and Li's *D* and *F* tests were applied to each dataset. The results of these tests are presented in Table 5.12., together with values calculated from the published *TaMS1* allele sequences. If a population is at neutral equilibrium these values should be close to zero. *TaMS1* displayed negative values for each of the three tests, although the exact significance of these results is difficult to interpret. Such low values may arise from several sources including recent population expansion, directional selection or an excess of low frequency mutants. Previous studies have highlighted the influence of sample design in generating artefactually low values of these indices (Ptak and Przeworski 2002; Hammer *et al.* 2003) which, for example, may be caused by pooling samples from different populations allowing fine-scale geographical differentiation to artificially increase the proportion of rare alleles and thus reduce the values. To counteract this effect, each test was also performed on the Turkish sequences alone for the six study genes. The C9 sequence was not included and the dataset comprised only the sequences generated from parasite DNA isolated from the four cattle in Sariköy village (Table 5.2.).

For *mero1* each test was performed on the datasets representing both the PCR error corrected sequences and the uncorrected sequences (Table 5.12.). For the uncorrected dataset, *mero1* showed a positive value of 0.269 for Tajima's *D* test across all sequences and a value of 0.316 across the Turkish samples. Although positive, and therefore suggestive of balancing selection, these results are not statistically significant. Coalescence simulations were used to determine a 95 % confidence interval for all three indices. If the calculated value lies within this interval then the null hypothesis of neutrality could not be rejected. Since the population analysis showed that *T. annulata* is panmictic, the model used to calculate this interval was based on free recombination. The results for *mero1* were below this limit for each test (Table 5.12.). When the PCR error corrected dataset were analysed, positive values were obtained for each of the three tests over all the allelic sequences and for the Turkish sequences alone. In the case of Tajima's *D* test and Fu and Li's *F* test, the values were statistically significant, because they exceeded the upper limit of the 95 % confidence interval. The difference between the 'corrected' and 'uncorrected' datasets was directly attributed to the removal of the putative PCR errors. Such errors would manifest themselves as truly neutral mutations within an allelic sequence and when present in sufficient quantity they are predicted to reduce the power of these tests to detect a departure from neutrality. The Turkish population may be considered the most suitable dataset, as the effect of geographical sub-structuring can be

Table 5.12. Tests for departure from neutrality

The null hypothesis that each of the allelic datasets represented neutrally evolving genes was tested using Tajima's *D* test and Fu and Li's *D* and *F* tests (Section 5.2.4.4). The 95 % confidence interval for each test was calculated by coalescence modelling, based on free recombination and the value of θ_π; a test result greater than the upper limit of this interval indicated a statistically significant departure from neutrality. Each test was performed on the entire dataset for each gene and also using sequences from Turkey alone (excluding the C9 sequence). This was done to avoid the influence of genetic differentiation between populations, which was highly significant for *SuAT*$_1$. Test values above the upper 95 % confidence limit are coloured in **red**. A sliding windows representation of the result of these three tests for *mero1* is presented in Figure 5.10.

Table 5.12. Tests for departure from neutrality

Gene name	Nucl.	Overall							Turkey						
		n	S	π	θ_π	*D* (Tajima)	*D* (Fu & Li)	*F* (Fu & Li)	n	S	π	θ_π	*D* (Tajima)	*D* (Fu & Li)	*F* (Fu & Li)
TaMS1	822	119	236	0.063	44	- 0.102	- 1.955	- 1.342	-	-	-	-	-	-	-
						0.358	*0.696*	*0.575*	-	-	-	-	-	-	-
mero1 *	654	58	122	0.049	26	0.269	-0.731	-0.411	47	113	0.048	26	0.316	-0.386	-0.151
						0.505	*0.676*	*0.692*					*0.478*	*0.658*	*0.671*
mero1 †	654	58	99	0.048	22	**0.878**	0.547	**0.808**	47	96	0.047	22	**0.811**	0.628	**0.831**
						0.513	*0.778*	*0.731*					*0.550*	*0.726*	*0.717*
mero2	1,461	11	118	0.025	39	- 0.673	- 0.663	- 0.759	7	87	0.023	35	-0.563	-0.479	-0.552
						0.372	*0.400*	*0.495*					*0.364*	*0.390*	*0.411*
SVSP1	1,494	23	151	0.028	41	- 0.192	- 0.947	- 0.834	12	111	0.028	37	0.389	0.011	0.126
						0.399	*0.449*	*0.468*					*0.393*	*0.387*	*0.442*
SVSP2	1,194	21	101	0.025	32	- 0.437	- 0.520	- 0.579	13	92	0.024	30	-0.370	-0.039	-0.147
						0.448	*0.522*	*0.565*					*0.452*	*0.464*	*0.498*
TashHN	996	17	20	0.006	6	0.155	-0.677	-0.510	10	13	0.005	5	0.137	-0.263	-0.184
						1.121	*1.044*	*1.139*					*1.031*	*0.961*	*1.119*
SuAT$_1$	1,359	16	100	0.031	30	**1.388**	**0.762**	**1.088**	7 ‡	17	0.003	5	-1.228	-0.789	-1.049
						0.619	*0.731*	*0.747*					*0.920*	*1.069*	*1.173*

The number in italics below the value for each test is the upper limit of the 95 % confidence interval, calculated by coalescence simulation based on θ_π and free-recombination, n = number of sequences compared, S = number of segregating sites, π = nucleotide polymorphism, θ_π = theta (an estimator of genetic variation) based on nucleotide polymorphism, * *mero1* dataset before PCR error correction, † *mero1* dataset following PCR error correction, ‡ Turkish-type allele

excluded within this collection of samples. Values of 0.811 and 0.831 were obtained for Tajima's *D* and Fu and Li's *F* test respectively with the value of Fu and Li's *D* test close to the upper threshold. Taken together these three indices provided supportive evidence of balancing selection influencing *mero1* diversity. To analyse this deviation from neutrality across the gene, a sliding window approach was taken using a 60 bp frame and a step size of 20 bp. The resultant plots of Tajima's *D* and Fu and Li's *D* and *F* tests are presented in Figure 5.10. For each test, a consistently positive value can be seen in the base range 220 – 500, which is above the 95 % confidence interval, providing clear evidence of balancing selection. This corresponded to the region of highest nucleotide diversity (Figure 5.6.). A low, increasing value for Tajima's *D* test and Fu and Li's *D* and *F* tests in the base range of 0 – 100 suggested that the signal peptide was not under the influence of balancing selection. Unexpectedly, the GPI-anchor encoding region of the gene displayed a positive value for Tajima's *D* test suggesting that it may not be diversifying neutrally. However, earlier $d_N d_S$ analysis suggested that this domain was conserved at the amino acid level.

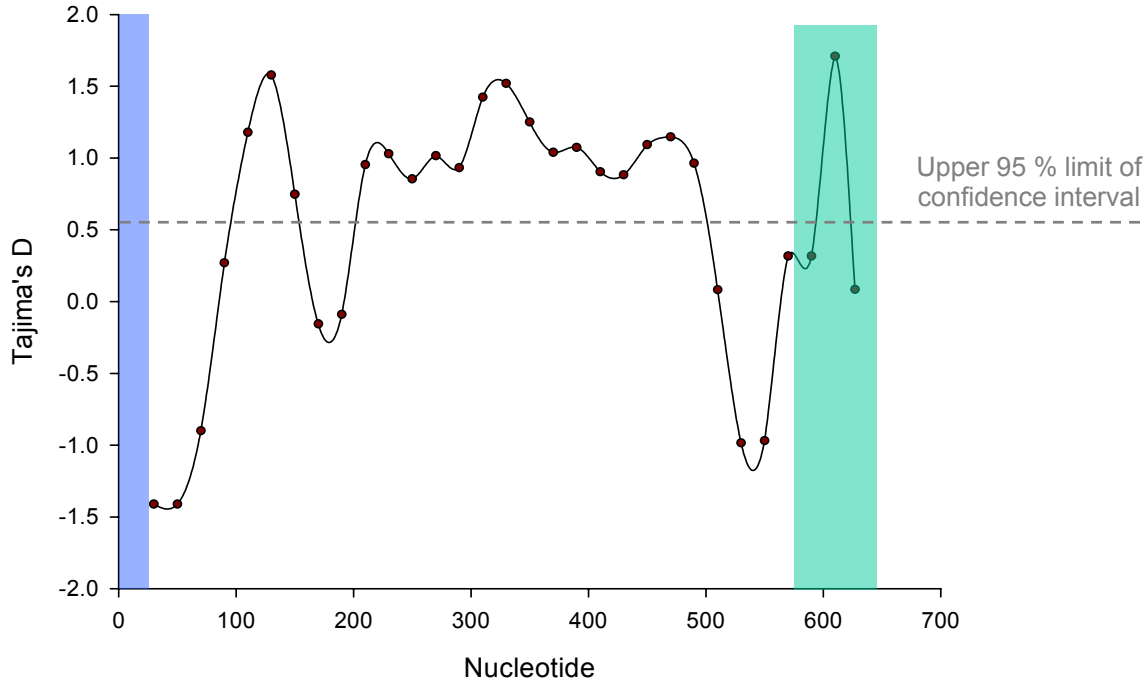
The *mero2* and *SVSP2* genes did not generate positive values for any of the three tests either in Turkey or over all the sequences (Table 5.12). Therefore, there was no evidence these genes departed from neutrality. *SVSP1* generated low positive values for each test using the Turkish sequences, although none of these were statistically significant. *TashHN* generated a low positive value for Tajima's *D* test overall and within the Turkish sequences. This was far below the relatively high threshold for statistical significance, which in turn was attributed to the very low level of polymorphism demonstrated across the alleles ($\pi = 0.006$, $S = 20$). The results of *SuAT1* underscore the value of comparing homologous alleles from a single population. When the sequences from Tunisia and Turkey were pooled and analysed, statistically significant positive values were generated for all three tests. In the case of Tajima's *D* test, the value of 1.388 was over double the 95 % confidence threshold of 0.619. This provided strong evidence of a departure from neutrality for this locus. However, when the seven Turkish-type alleles were analysed independently, the values for each test dropped below zero. A departure from neutrality was indicated using the entire sequence set because different allelic types were compared. As previously discussed in Section 5.3.5., cluster analysis demonstrated significant geographical sub-structuring between Tunisian and Turkish populations at the *SuAT1* locus (Figure 5.8.(vi)). Therefore, it was known *a priori* that all the alleles could not be considered as diversifying neutrally and the results were considered consistent with sub-structuring and maintenance of different allelic types.

Figure 5.10. Neutrality tests on Turkish sequences of *mero1*

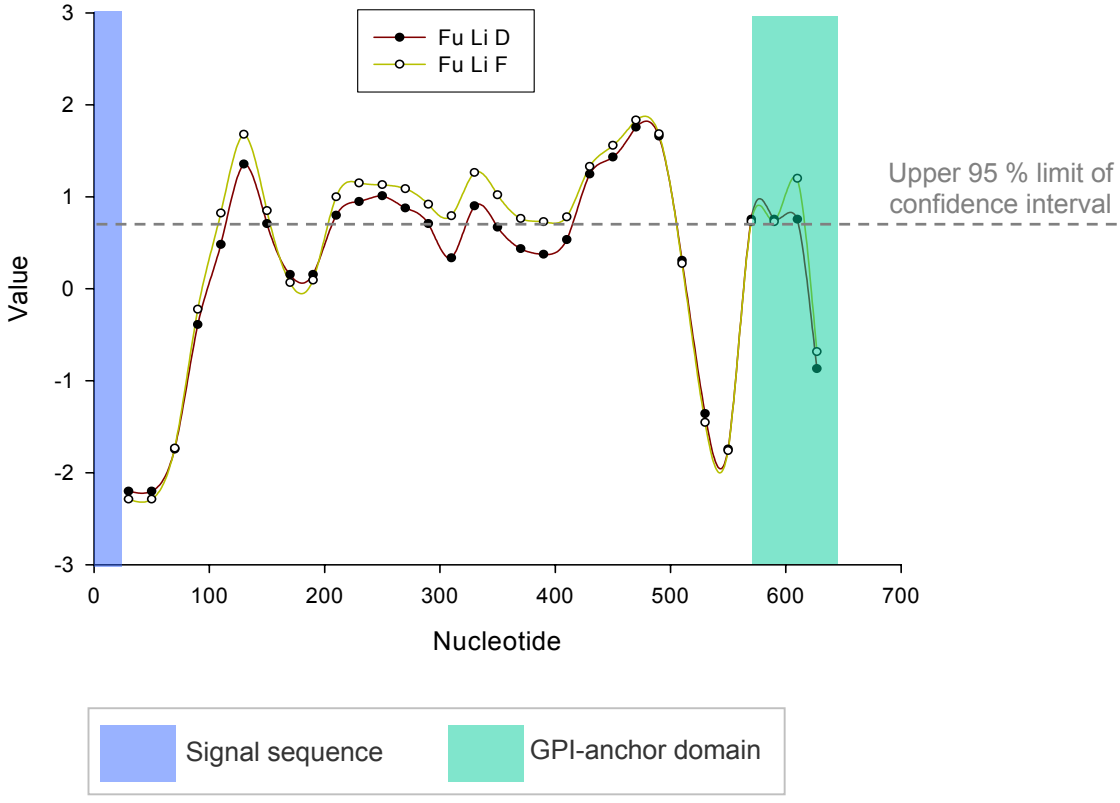
Three statistical tests of neutrality were performed on 46 sequences, corresponding to Turkish alleles of the *mero1* gene. The results of Tajima's *D* test are presented in **(i)**, while results of the related Fu and Li's *D* and Fu and Li's *F* tests are presented in **(ii)**. The dataset used in these analyses was corrected for PCR errors by the method described in section 5.4.4. To illustrate the location of various peptide motifs in the translated gene product, a graphic of the gene is also presented.

Figure 5.10. Neutrality tests on Turkish sequences of *mero1*

(i) Tajima's *D* test



(ii) Fu and Li's *D* and *F* tests



5.3.8. Polymorphism in TashAT family proteins

The analysis of the two TashAT gene family members, *SuAT$_1$* and *TashHN*, showed contrasting patterns of variation although, in principal, their functions may be very similar. In order to provide a context for explaining these results, it is necessary to briefly describe the TashAT locus.

At the TashAT locus, the flanking members *TashAT$_2$*, *TashHN*, TA03115 and TA03110 have orthologous genes identified in the flanking members of the family in *T. parva* (Figure 5.11.). The amino acid sequences of the sixteen genes in *T. annulata* and twenty genes in *T. parva* were aligned, clustered and used to generate a dendrogram of sequence similarity (Figure 5.12.). Notably, the four flanking genes clustered with their *T. parva* orthologues, whereas the majority of the internal genes formed two species-specific clusters. This suggested that the flanking genes may be ancestral and that intervening genes, in general, could be the product of tandem gene duplication in each species. When the sixteen family members in the C9 sequence of *T. annulata* were aligned, a conserved motif, `L(Q/E)PETIPVE` was identified across the eleven paralogues (Figure 5.13.). Two copies of this motif were found in SuAT$_1$ – the first at amino acid position 352 – 360 (Figure 5.13.) and a second at amino acid position 494 – 502 (not shown). Although the function of this motif is currently unknown, its presence as multiple copies both within and between family members was striking. TashHN was one of the few family members that did not contain this motif.

When the sequences of the full-length alleles of TashHN from *T. annulata* were aligned, a four amino acid insertion / deletion was evident at the C-terminus of the protein (Figure 5.14.). Four amino acids (`TDTQ`) are present in the genome sequence of *T. parva* at this locus, while in *T. annulata* the sequence is `TESQ`, when present. The two substitutions (`D/E` and `T/S`) represent amino acids that possess similar biochemical properties. To assess the distribution of conserved and polymorphic nucleotides across the species, nucleotide diversity was calculated over the length of the un-gapped sequence and is shown in Figure 5.15. Comparison of the *T. parva* and *T. annulata* genome sequences shows 228 out of the 996 bases are different, giving an overall nucleotide diversity of 0.229. Although the 3' region showed the highest level of diversity, two highly conserved regions were identified at around 550 bp and 800 bp, located in regions annotated as PEST domains, suggesting the presence of a functional domain. The first conserved region begins with peptide sequence `KRRKYV` (not shown on alignment figure) and corresponds to a nuclear localisation signal (NLS) previously identified in TashHN (Swan *et al.* 2003).

Figure 5.11. Synteny between *T. annulata* and *T. parva* at the TashAT locus

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

Figure 5.11. Synteny between *T. annulata* and *T. parva* at the TashAT locus

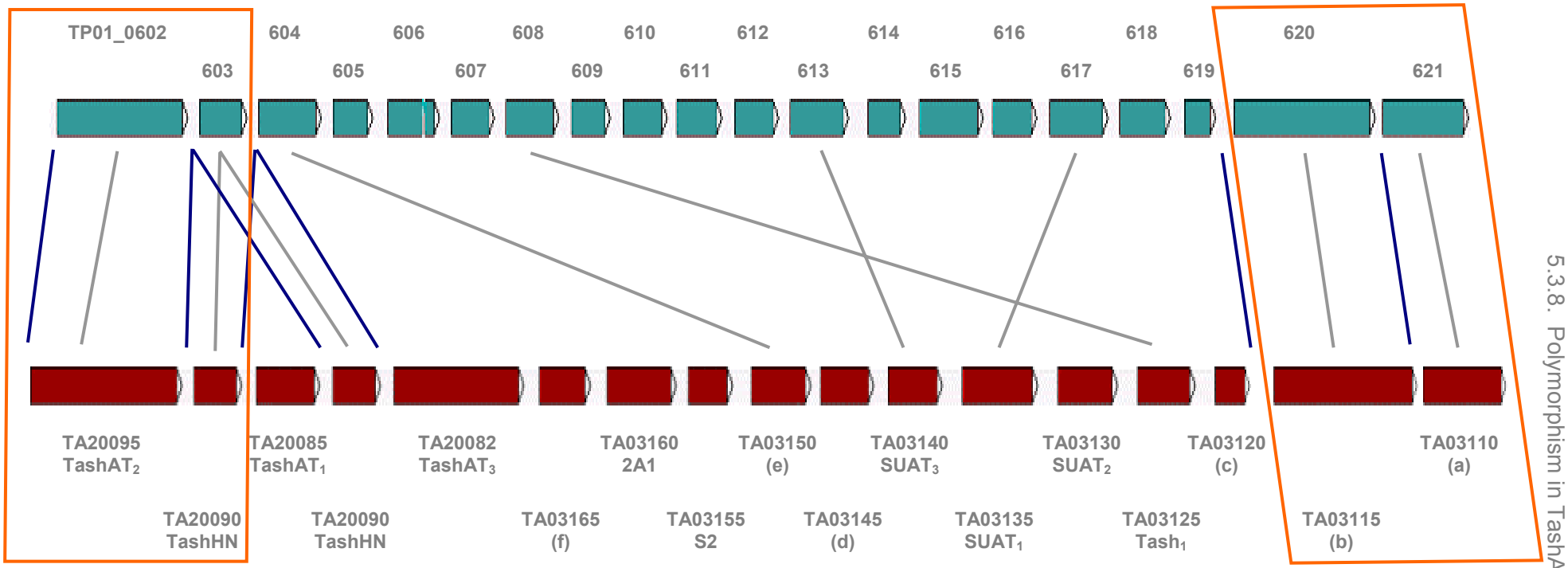


Figure 5.12. Dendrogram representing the TashAT family of *T. annulata* and orthologues in *T. parva*

The amino acid sequences of the sixteen genes in the TashAT family of *T. annulata* along with the twenty genes in the orthologous family in *T. parva* were aligned, clustered and used to create a dendrogram. *T. annulata* genes are highlighted in red while *T. parva* genes are highlighted in blue. The orthologous genes corresponding to each end of the TashAT locus (Figure 5.11.) are highlighted in a grey box. The presence of two discrete clusters of internal genes suggests that these genes are more closely related to each other within a species and have evolved independently in each species. This is supported by the fact that most internal genes do not have direct orthologues (Figure 5.11.).

Figure 5.12. Dendrogram representing the TashAT family of *T. annulata* and orthologues in *T. parva*

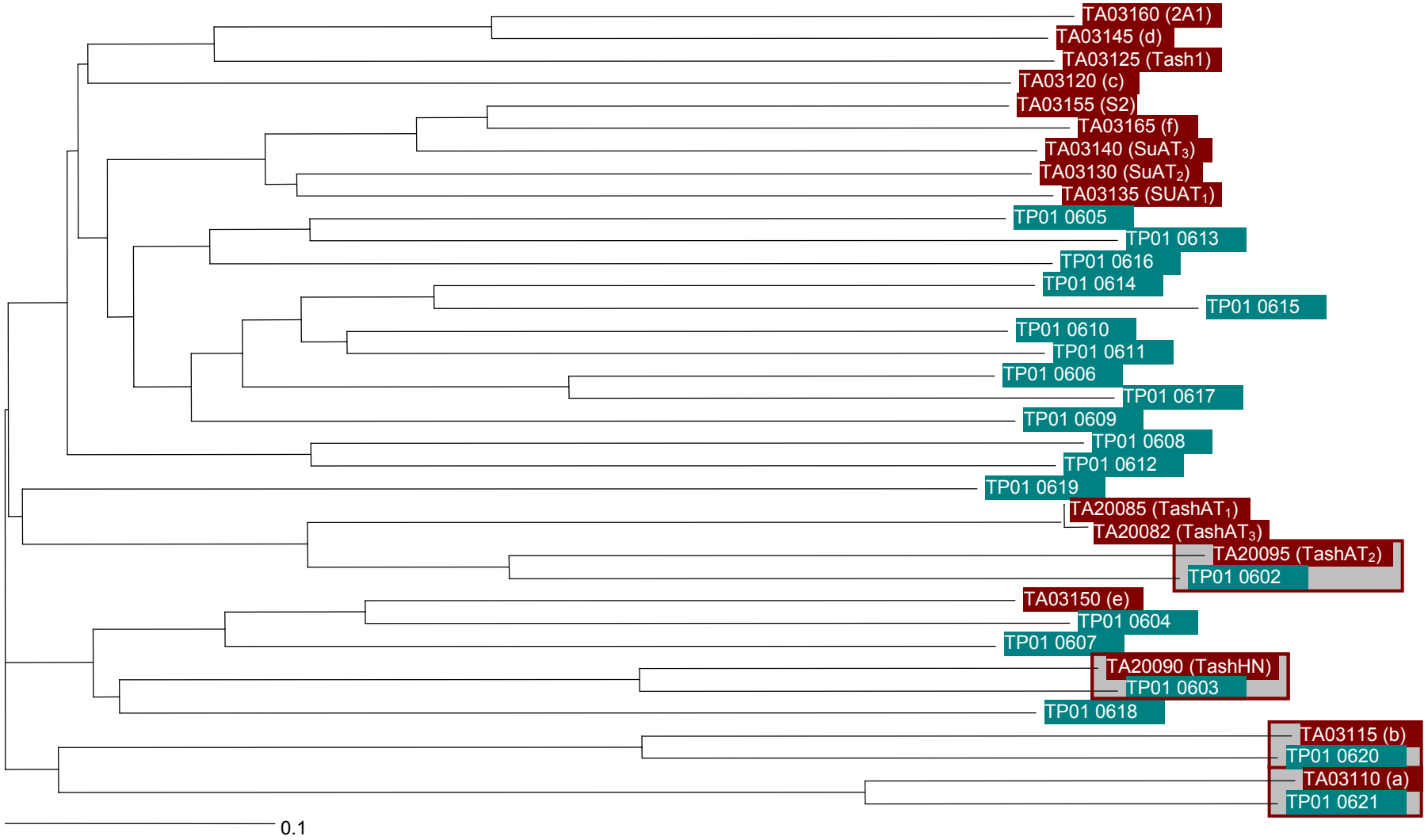


Figure 5.13. Conserved motif in TashAT family

The L(Q/E)PETIPVE motif is conserved among members of the TashAT family in the Ankara C9 sequence. Two excerpts of the amino acid alignment for several TashAT family members illustrates that the motif is double copy in the Tash 1 gene. A second copy of the motif is also found in $SuAT_1$, although this does not appear in either of these multiple alignments due to the extensive polymorphism between family members in other regions. TashHN was found not to possess this motif. Significantly, this motif is conserved in the hypervariable region of SVSP1 among both Tunisian and Turkish alleles (Figure 5.17.).

Figure 5.13. Conserved motif in TashAT family

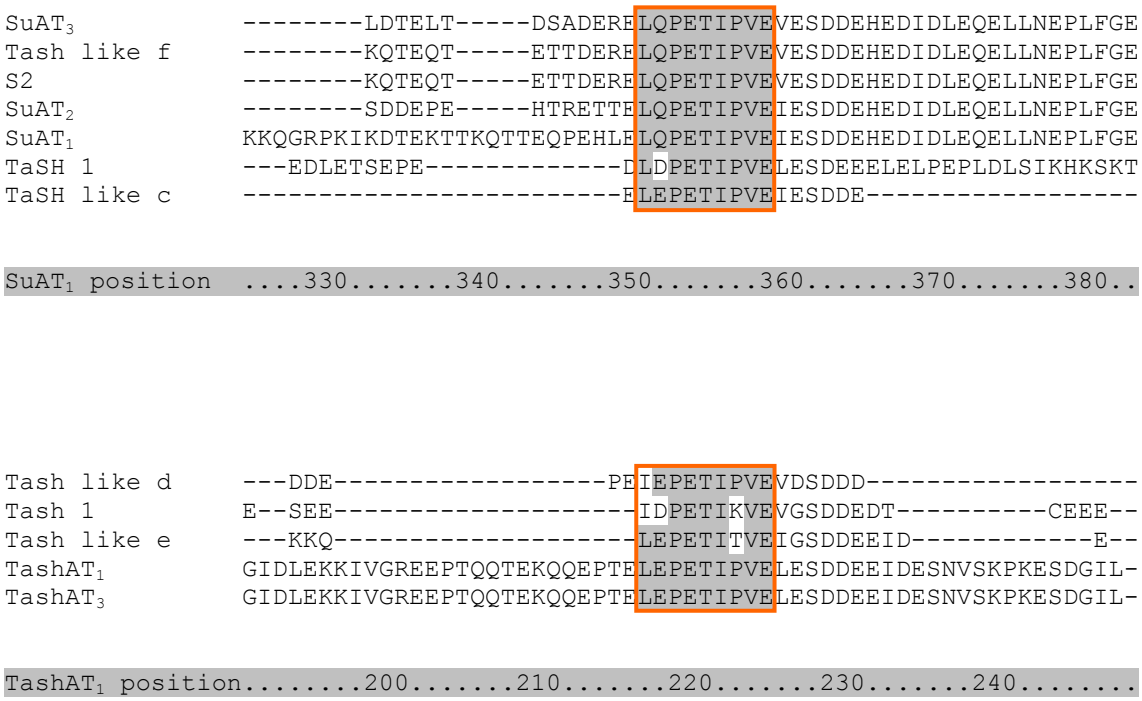


Figure 5.14. Carboxyl terminal of TashHN alleles and TpshHN

Amino acid sequences representing alleles of TashHN, together with a single allele of its orthologue in *T. parva*, TpshHN were aligned using ClustalX. The C-terminal region of the alignment is presented opposite. At each site the second most frequent amino acids are indicated.

Figure 5.14. Carboxyl terminal of TashHN alleles and TpshHN

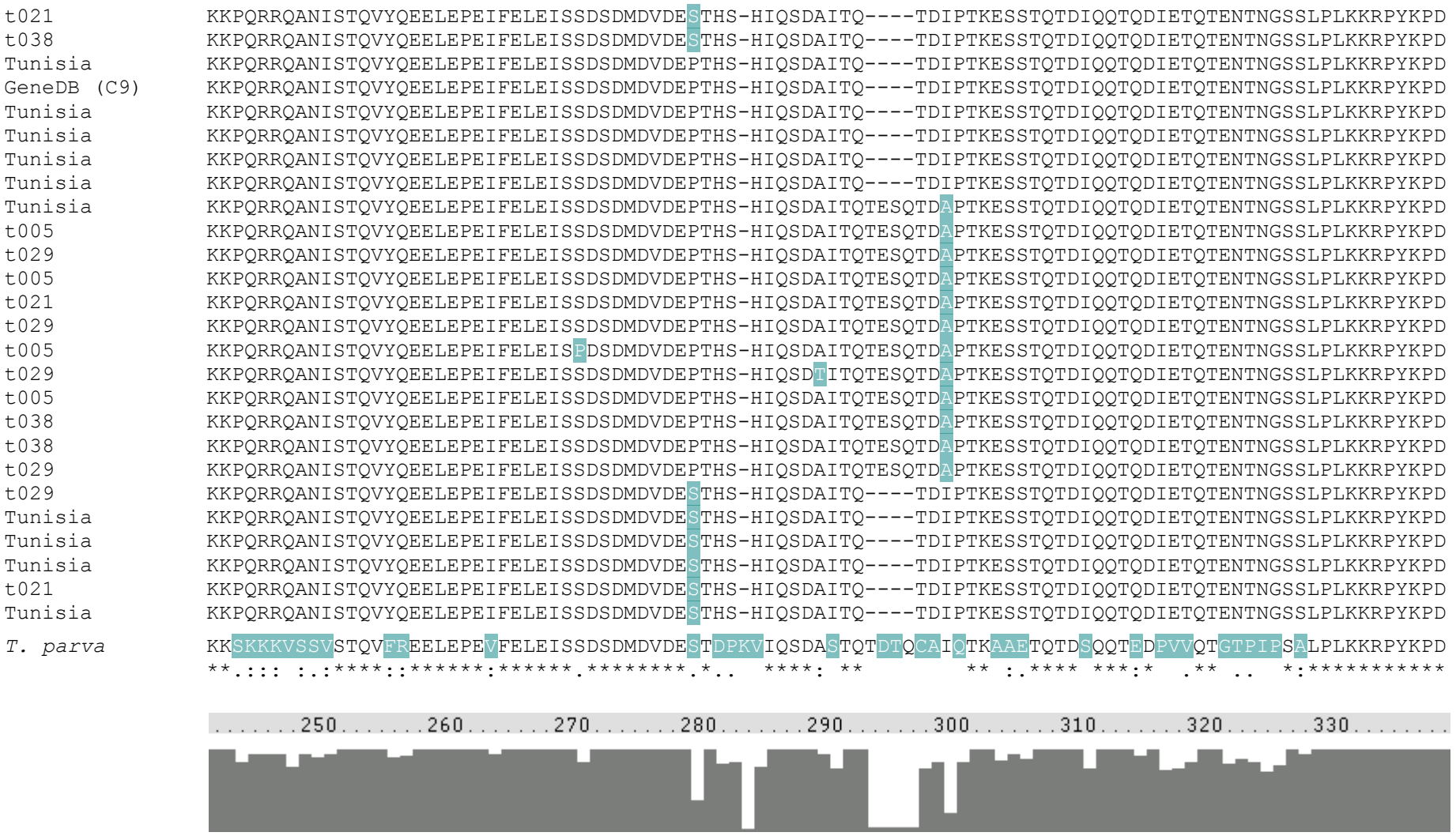
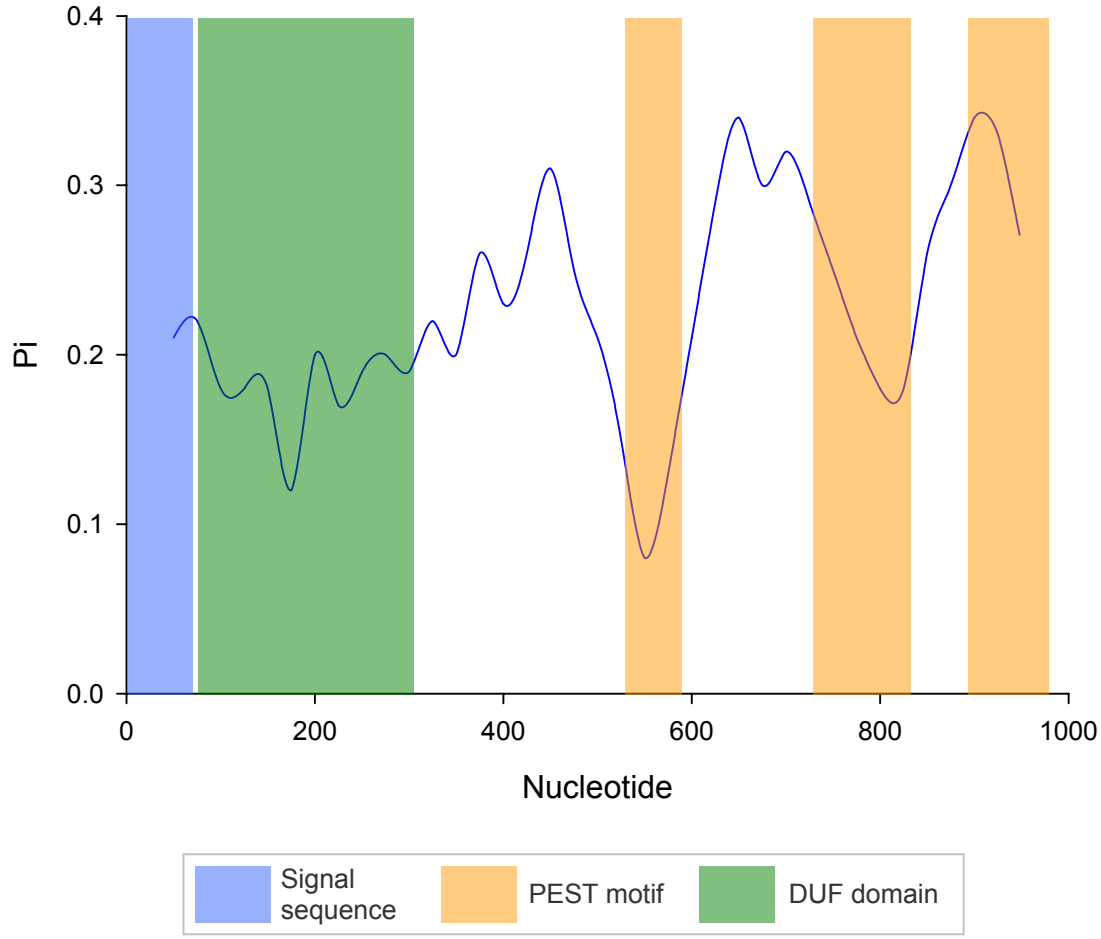


Figure 5.15. Nucleotide diversity between *TashHN* (C9) and *TpshHN*

The C9 nucleotide sequence of *TashHN* was compared to that of its orthologue in the genome of *T. parva*, *TpshHN*. Following sequence alignment and gap removal, nucleotide diversity was calculated across the length of gene. The position of peptide motifs in the *T. annulata* sequence is highlighted.

Figure 5.15. Nucleotide diversity between *TashHN* (C9) and *TpshHN*



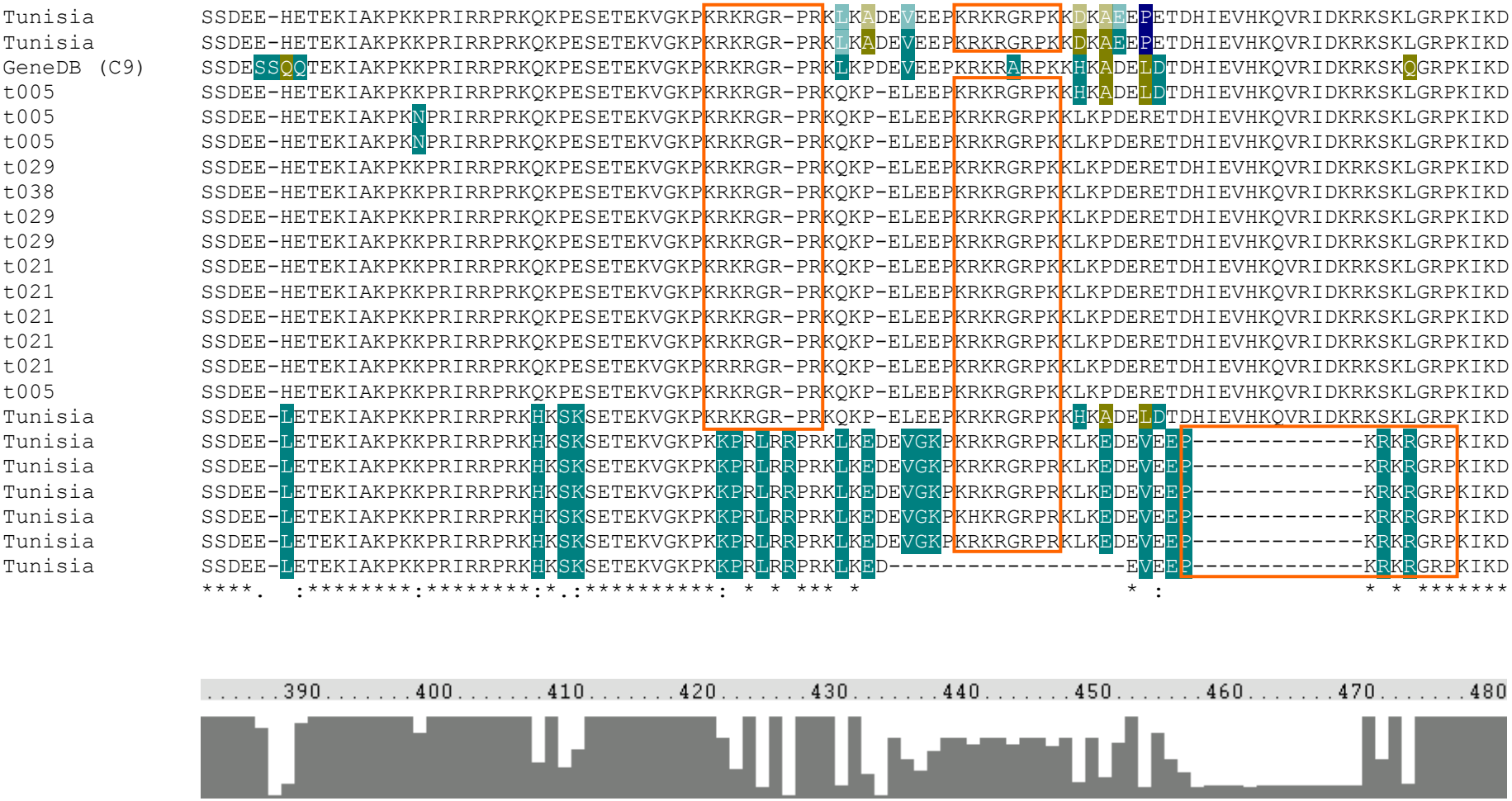
The second region is a PEST domain just preceding the insertion / deletion site (Figure 5.14., position 259 - 280). Additionally, the eleven amino acids at the C-terminus are completely conserved within *T. annulata* alleles and between *T. parva* and *T. annulata* and a NLS has also been identified in this area – PLKKRPY. The results indicate that two NLS motifs, in particular, are highly conserved in *T. annulata* and are also conserved in *T. parva*. It has been shown experimentally that TashHN locates to the host cell nucleus (Swan *et al.* 2003) and consequently these results suggest that this activity is critical for the parasite.

The results of the $SuAT_1$ gene presented the greatest challenge in terms of interpretation, although it was possible to reconcile evidence from different analyses. $SuAT_1$ is an internal family member and is orthologous to TP01_0617 in *T. parva* (Figure 5.11.), although the *T. parva* orthologue encodes only 426 amino acids compared to 558 in the C9 strain of *T. annulata*. This represents a 3' terminal extension of the gene in *T. annulata*. Amino acid identity between the orthologues is only 47 %, a level comparable with the SVSP proteins, although the pattern of allelic diversity was quite distinct from the latter genes. It was notable that different sets of PCR primers were required to amplify the samples from each country (Figure 5.2.), suggesting a geographical divergence in sequence between the two regions and a lack of conservation of priming sites. This was supported when the gene was shown to exhibit a low level of nucleotide diversity in each population with a large difference becoming apparent when populations were pooled (Table 5.9.). The McDonald-Kreitman test indicated a departure from neutrality (Table 5.10.) and when positive and negative selection was charted across the gene, an excess of positive substitutions was evident, Figure 5.9.(vi), although none of these were statistically significant. A large number of synonymous and non-synonymous substitutions can be seen between codon 400 and 450. It must be borne in mind that almost 20 % of the coding sequence was not analysed from the 3′ terminus of the gene. This represented the small portion of the gene that was not sequenced and also included gaps and local areas of block misalignment. Several sites were identified where insertion / deletions occurred in the multiple alignment, which were located within a polymorphic region towards the C-terminus of the protein (Figure 5.16.). The segregation of sequences into Tunisian-type, Turkish-type and an intermediate form is evident and corresponds with the cluster analysis (Figure 5.8.(vi).). A block deletion in the Tunisian-type alleles at amino acid site 459 – 471 contrasts with complete conservation in the Turkish sequences (Figure 5.16.). Two Tunisian sequences and two Turkish sequences (including C9) represent intermediate allelic types and are positioned at the top of the multiple alignment. Conserved AT hook DNA binding motifs

Figure 5.16. DNA binding motifs in $SuAT_1$

Amino acid sequences representing alleles of $SuAT_1$ were aligned using ClustalX. At each site the second, third, fourth most frequent amino acids are indicated while DNA binding motifs termed 'AT hooks ' are highlighted in orange. Block differences distinguish the seven Tunisian alleles at the bottom of the alignment and the Turkish derived alleles (t005, t021, t029, t038). The two Tunisian sequences and the C9 sequence at the top of the alignment are intermediate between the archetypal Tunisian and Turkish alleles.

Figure 5.16. DNA binding motifs in $SuAT_1$



were found in both the Tunisian-type, Turkish-type and intermediate-type alleles. Similar to the nuclear localisation sequences in *TashHN*, the conservation of these motifs in *SuAT1* is indicative that they are functionally important domains.

Several differences between the Tunisian and Turkish-type sequences of *SuAT1* can be seen in Figure 5.16. at positions 409 – 412. This area is within consensus positions 400 – 450 in Figure 5.9.(vi) where an abundance of positively and negatively selected sites were found and it may be argued that this constitutes an area of amino acid misalignment. Other, more obvious examples of block differences between Tunisian and Turkish sequences were found in the centre of the gene (data not shown) and were excluded from the $d_N d_S$ and other analyses at the outset of the study. The presence of such well-conserved blocks is consistent with different allelic types evolving in isolation, i.e. geographical sub-structuring. The forced alignment of remaining small variant blocks of sequence is likely to result in an overestimation of non-synonymous substitutions and may explain the high global $d_N d_S$ value for this molecule. To gain more insight into the underlying pattern of diversity of this gene, the Tunisian and Turkish samples should be separated when performing the McDonald-Kreitman test and $d_N d_S$ analysis with the small number of putatively misaligned regions removed from the multiple alignment. This would entail re-analysing smaller portions of a limited number of sequences. This was not performed because (1) the dataset would become quite small, (2) the reduced amount of alignment would not represent the gene as a whole as was the intention of the study and (3) it was beyond the scope of the project, since the host nuclear genes were intended for their comparative value against the merozoite candidates and the SVSP proteins. However, to exclude the influence of misalignment between allelic types, for the tests of neutrality (Table 5.12.), the Turkish-types sequences were analysed independently with the results indicating that *SuAT1* showed similar diversity to *TashHN* and did not depart from neutrality.

Some regions of *SuAT1* were completely conserved between Tunisian, Turkish and intermediate-type alleles. This includes the L(Q)PETIPVE motif, which is commonly found in TashAT family genes and occurs twice in *SuAT1*. The second copy of this motif lay outside the sequenced region of the gene, however the first copy, at position 352 – 360, was completely conserved as LVPETIPVE. Although the function of this motif is unknown, like the NLS and AT hooks, the fact that it is completely conserved implies it has functional significance. A very similar motif was found to be conserved in the SVSP genes.

5.3.9. Polymorphism in SVSP family proteins

The SVSP genes chosen for this study demonstrated a large amount of polymorphism at both the DNA and amino acid level. *SVSP1* contains three regions where an insertion / deletion event has taken place (see Figure 5.5.). Two of these are short and invariant and are located at either end of the coding sequence. The third region represents an extensive, 180 bp hypervariable region in the centre of the molecule and corresponds to a large PEST domain. The aligned amino acid sequences of allelic variants in this hypervariable domain can be seen in Figure 5.17. In the central part of this region a conserved peptide motif can be distinguished – L(Q/*)PET(V/I)PVE(V/I) followed by conserved residues G and S. This is homologous to the LQPETIPVE motif identified in the TashAT family and conserved across all the alleles of $SuAT_1$. Coincidentally, a BLASTP search of the *T. annulata* genome database using TaSP as a query revealed a degree of similarity to a large number of SVSP members (Figure 5.18.) with similarity limited to a motif in the centre of the molecule - LQPETVTVEV. Ten of the top twelve hits (down to an expected 'cut-off' value of 1.1×10^{-5}) were SVSP proteins. A previous study on allelic sequences of TaSP revealed this area was relatively well conserved at the allelic level with only three dimorphic sites L(Q/E)PE(T/S)V(T/S)VEV (Schnittger *et al.* 2002), although it is flanked by highly variable regions. Similar to TaSP, there are two low complexity regions rich in glutamic acid (E), proline (P), threonine (T) and glutamine (Q) with some serine (S) residues, flanking this motif in *SVSP1,* which constitute PEST domains. When *SVSP1* is so polymorphic across its length, why is this region particularly diverse? Due to misalignment and complex sequence gaps, it was impossible to conduct $d_N d_S$ and McDonald-Kreitman analyses over this region, so it has not been formally determined whether the diversity is consistent with positive selection. However, there must be some underlying reason which can explain this high level of polymorphism. Nevertheless, the presence of the conserved motif between SVSP1, the TashAT family and TaSP suggests these all belong to a larger family of proteins.

The PEST domain in SVSP2 had much fewer gaps in the amino acid alignment of allelic sequences (Figures 5.5.(ii)). The larger of these two regions in SVSP2 is presented in Figure 5.19., where identical sequences were derived from all four Turkish isolates and several Tunisian clones. Varying sizes of block substitution can be seen at the bottom of the alignment, including a Tunisian and Turkish (t005) sequence, which bear little resemblance to the predominant sequence at the locus. In general, SVSP2 contained few alignment gaps and consequently only a small proportion of the gene was excluded from

Figure 5.17. Hypervariable region of SVSP1

Amino acid sequences representing alleles of SVSP1 were aligned using ClustalX, identifying hypervariability in the PEST domain in the centre of the protein. At each site the second, third, fourth and fifth most frequent amino acids are indicated and the conserved motif L(Q/*)PET(V/I)PVE(V/I) is highlighted in orange. The graphic beneath the sequence alignment summarises diversity at each amino acid site.

Figure 5.17. Hypervariable region of SVSP1

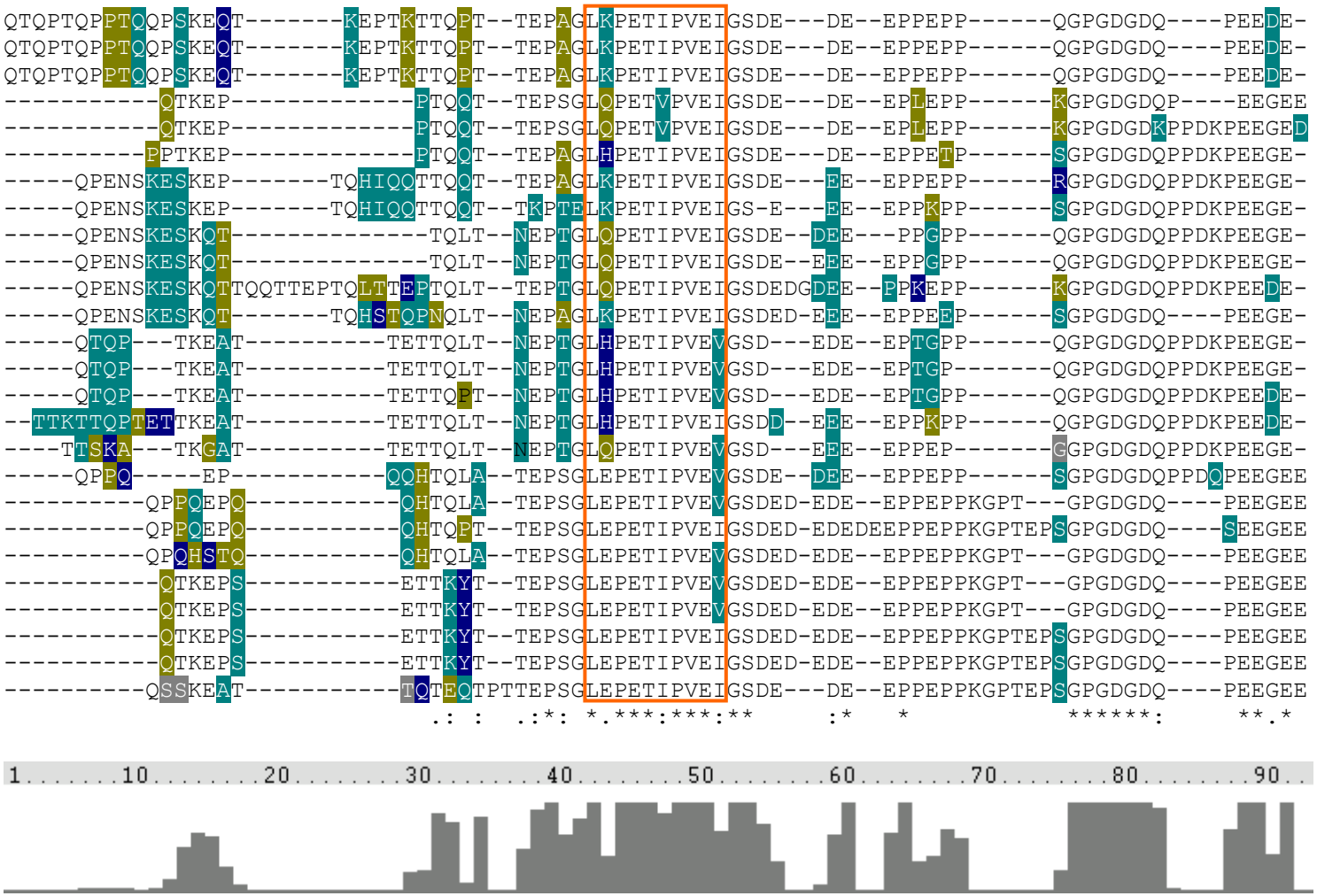


Figure 5.18. TaSP BLASTP search results

The amino acid sequence of the macroschizont surface protein, TaSP, was used as a query to perform a BLASTP search of the coding sequences of the *T. annulata* genome. The results of this search are presented opposite. Sub-telomeric variable secreted protein (SVSP) family members, including SVSP1 (TA16025), were identified as high-scoring matches. The amino acid similarity between SVSP1 and TaSP is also presented along with a schematic diagram illustrating the location of peptide motifs and regions of significant allelic polymorphism in the SVSP1 gene.

Figure 5.18. TaSP BLASTP search results

```
Low complexity filtering disabled
Repeatmasker disabled

BLASTP 2.0MP-WashU [16-Sep-2002] [decunix4.0-ev6-I32LPF64 2002-09-18T19:28:12]

Copyright (C) 1996-2002 Washington University, Saint Louis, Missouri USA.
All Rights Reserved.

Reference:  Gish, W. (1996-2002) http://blast.wustl.edu

Query=  TaSP protein (314 letters)

Database:  GeneDB_Tannulata_Proteins
           3795 sequences; 2,026,930 total letters.
```

Sequences producing High-scoring Segment Pairs:	High Score	Smallest Sum Probability P(N)	N
TA17315 |||surface protein precursor (TaSP)|Theileria ann...	1694	2.3e-176	1
TA17535 |||Theileria-specific sub-telomeric protein, SVSP...	136	7.2e-08	1
TA17120 |||Theileria-specific sub-telomeric protein, SVSP...	131	2.3e-07	1
TA09800 |||Theileria-specific sub-telomeric protein, SVSP...	129	3.8e-07	1
TA18860 |||conserved Theileria-specific sub-telomeric pro...	128	6.1e-07	1
TA20980 |||proline-rich hypothetical protein|Theileria an...	126	1.2e-06	1
TA09790 |||Theileria-specific sub-telomeric protein, SVSP...	123	1.6e-06	1
TA05545 |||Theileria-specific sub-telomeric protein, SVSP...	119	2.3e-06	1
TA17346 |||Theileria-specific hypothetical telomeric sfii...	117	2.9e-06	1
TA09785 |||Theileria-specific sub-telomeric protein, SVSP...	117	6.2e-06	1
TA16040 |||Theileria-specific sub-telomeric protein, SVSP...	117	6.4e-06	1
TA09805 |||Theileria-specific sub-telomeric protein, SVSP...	115	9.3e-06	1
TA11385 |||Theileria-specific sub-telomeric protein, SVSP...	116	1.1e-05	1
. . . . .			
. . . . .			
. . . . .			
TA16025 |||Theileria-specific sub-telomeric protein, SVSP...	90	0.0055	1

```
>TA16025 Theileria-specific sub-telomeric protein, SVSP family
        Length = 560

 Score = 90 (36.7 bits), Expect = 0.0056, P = 0.0055
 Identities = 25/68 (36%), Positives = 35/68 (51%)

Query:    57 QPAQQEPIEPQQPTQPSTEPEELQPETVTVEVPEPVTSEEPKESDQTEEQKHEEPEASPA-PEPVDEP
123
             +P +Q     QQ TQ +TEP  L+PET+ VEV           SD+ E+++  EP   P  P   D+P
Subject: 275 KPTKQPQHSTQQHTQLATEPSGLEPETIPVEVG----------SDEDEDEEPPEPPKGPTGPGDGDQP
332
```

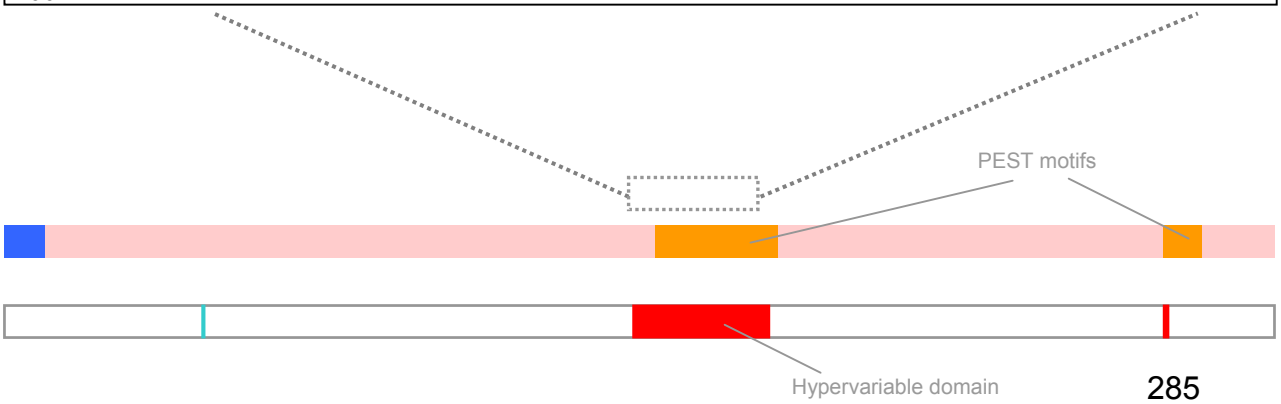
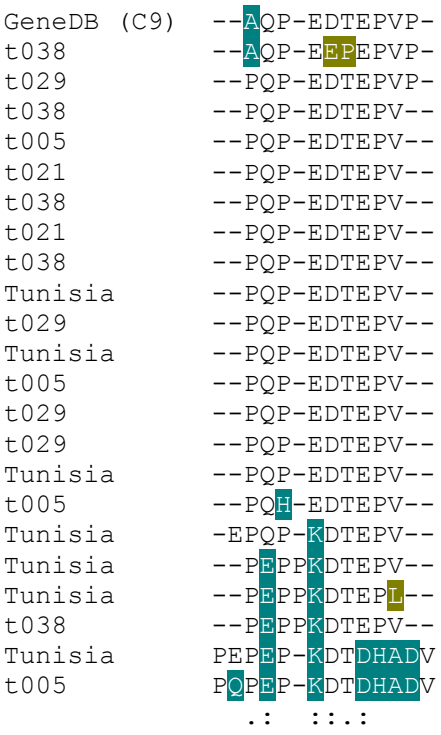


Figure 5.19. Region of SVSP2 containing alignment gaps

Amino acid sequences representing alleles of SVSP2 were aligned using ClustalX. At each site the second, third and fourth most frequent amino acids are indicated. The small PEST domain in this protein has relatively few gaps in the amino acid alignment compared to the hypervariable PEST domain in SVSP1 (Figure 5.17.).

Figure 5.19. Region of SVSP2 containing alignment gaps



the analysis, with the majority of the PEST sequence being incorporated in McDonald-Kreitman and $d_N d_S$ tests. The highest level of nucleotide diversity can be seen within this region of *SVSP2* (Figure 5.6., around nucleotide 400) in both the Tunisian and Turkish populations, which is echoed in the $d_N d_S$ plot in Figure 5.9.(iv). In contrast, the remainder of this PEST motif (following removal of gaps in the alignment) is under purifying selection, indicated by low d_N minus d_S values.

The SVSP genes are clearly highly variable, with *SVSP1* displaying the greatest intra-species $d_N d_S$ value of all the genes in this study. Interestingly SVSP genes showed more sites under positive selection and fewer sites under negative selection at $p < 0.25$ compared to *mero1*, the gene proposed as under diversifying selection from the host immune system. Does this mean that SVSP genes are under more pressure to diversify than *mero1*? The most informative results probably came from the McDonald-Kreitman test, which returned a neutrality index of approximately one for both of these genes (Table 5.10.). This means there is an absence of intra-specific non-synonymous substitutions in *T. annulata* when compared with *T. parva* and the raw figures for synonymous and non-synonymous polymorphic differences are consistent with neutrality. The apparent paradox of this neutrally evolving, yet highly diverse gene family will be considered in the discussion.

5.4. Discussion

5.4.1. Summary of findings

Allelic sequence analysis of representative members of *T. annulata* gene families revealed several contrasting patterns of diversity. Two putative merozoite antigen genes displayed extensive polymorphism at the DNA level. Amino acid diversity and evidence of positive selection indicated that *mero1* may encode a merozoite surface antigen, which is under the diversifying selection of the bovine immune system. In contrast, the protein product of *mero2* was shown to be highly conserved casting doubts on its candidacy as an antigen gene. SVSP family genes were shown to be highly polymorphic both at the nucleotide and amino acid level, although there was no evidence of positive selection acting on either of the genes studied. Although not selected to diversify, the genes did not appear to be constrained by purifying selection and therefore it was proposed they exhibited a virtually neutral pattern of diversity. The host nuclear genes displayed contrasting patterns of diversity. *TashHN* was conserved at both the nucleotide and amino acid level, while $SuAT_1$ was highly polymorphic, particularly at the amino acid level with sequence analysis revealing several allelic 'types' at this locus. However, when Tunisian-type and Turkish-

type alleles of *SuAT1* were analysed independently, limited within-type nucleotide and amino acid diversity was observed. In contrast to the other genes studied, both host nuclear genes show marked evidence of geographical sub-structuring.

5.4.2. Genetic sub-structuring

Cluster analysis identified a degree of geographical sub-structuring between Tunisian and Turkish sequences with several of the genes under study, however complete separation between populations was never observed. The strongest evidence of sub-structuring was seen with the genes *SuAT1* and *TashHN*, which have host nuclear localisation signals. Geographical sub-structuring was less pronounced with *mero1* and *SVSP1* and completely absent with the *mero2* and *SVSP2* genes. The micro- and mini-satellite results presented in Chapters Two and Three already indicated some genetic differentiation between these areas. MLGs representing ten different loci were used to demonstrate that Tunisian and Turkish genotypes clustered separately (Figure 3.8.). This was supported by linkage analysis since elevated linkage disequilibrium was demonstrated when populations from Tunisia and Turkey were combined. However, none of micro- and mini-satellite loci alleles segregated between the countries, and it was necessary to combine the data from all ten loci over 305 samples in order to reveal clustering. Moreover, the initial marker study (Chapter Two) failed to show sub-structuring using these loci with a smaller population sample, until complete allelic profiles were analysed (Figure 2.11.). Why, then is it possible to cluster samples from each region using selected single loci identified from the sequencing study? This may be explained by several hypotheses, which are not mutually exclusive –

1. Quantity and quality of polymorphism. *SuAT1* and *TashHN* show limited nucleotide diversity when each population in considered in isolation. In contrast, micro- and mini-satellite markers are much more polymorphic, therefore a high level of diversity is observed in a particular locality, with a large number of alleles identified for each locus in *T. annulata*. Since only length polymorphism is taken into account when using this method of genotyping, sequence diversity at these loci is effectively discarded. Furthermore, mutation models to describe the diversity at each locus have not been identified; clustering was achieved by calculating proportion of alleles that are shared between any two individuals. If the length of an allele at a particular locus matches between two individuals then the cluster analysis designates 100 % similarity (even though there may be sequence diversity). However if the alleles are of a different length they are designated as 0 % similar, even though they may be phylogenetically very closely related.

The methodology for micro- and mini-satellite genotyping and MLG clustering may therefore have reduced power to accurately group individuals. The power of the technique rests on genotyping with a large number or markers, and for this purpose, ten loci may be too few. In contrast, the clustering algorithm used in the multiple sequence alignment procedure considers the full nucleotide sequence to group similar alleles. DNA sequences may be regarded as a richer dataset, where a gradation of diversity can be more readily observed and quantified. Furthermore, markers with a low number of alleles that are relatively abundant are generally accepted to be optimum for conducting population studies. Perhaps the *SuAT*$_1$ and *TashHN* genes that display limited diversity are more suitable for accurately gauging genetic differentiation between populations. However, this may seem unlikely, as the nature of their predicted gene products would indicate that they are unlikely to be neutral, with selective pressure(s) influencing their evolution.

2. Directional selection. It is possible that adaptive Darwinian selection is occurring in each of the two populations. Neutral micro- and mini-satellite loci however would be unaffected by such selection, and only the processes of genetic drift would lead to observable differentiation using these markers. In contrast, the results of selection are more noticeable on the parasite encoded host nuclear genes and to a lesser extent on *mero1* and *SVSP1*. Could it be that directional selection is particularly acting on genes that may operate to modulate host (cattle or tick) cell phenotype within *T. annulata* and if so, what differences in the ecology / biology of the parasite between each country could drive these genes to diverge? The first possible difference to consider is local variation in cattle breed. However, it is difficult to imagine that differences in cattle breeds are so profound to account for such a mechanism. As discussed in Section 1.7.1., the parasite is transmitted in Tunisia solely by the two-host tick, *H. detritum*. In Turkey, in addition to *H. detritum*, three different three-host *Hyalomma* species are responsible for disease transmission. If, indeed, different tick species correlated with different allelic types, directional selection would be consistent with host (i.e. vector) sub-structuring. This may be investigated in future studies by genotyping *T. annulata* isolated from different species of ticks from within a single locality in Turkey.

3. Artefact of sampling method. Sampling from four cattle from a single site in Turkey and comparing them to ten isolates originating in scattered locations throughout Tunisia is not an ideal sampling strategy for demonstrating and explaining geographical sub-structuring. Inbreeding within the Turkish sample is highly likely, with frequent recombination of local genotypes of *T. annulata* occurring at a high rate. This coupled

with the fact that a degree of sub-structuring had already been identified between the countries means it is perhaps not surprising that a proportion of genes display allelic diversity which correlates with their origin. Additionally, there is the possibility that selective pressures on the TashAT family genes make them susceptible to this type of influence. However, it must be remembered that the primary aim of this study was simply to identify sufficient polymorphism in comparable alleles to conduct statistical tests to detect signatures of positive selection.

5.4.3. Suitability of sampling strategies

The two sampling strategies were designed without *a priori* knowledge of the level of polymorphism that would be identified in each gene. Two of the genes (*mero1* and *SVSP1*) showed greater nucleotide diversity in the Turkish population while *SVSP2* and $SuAT_1$ showed more diversity within the Tunisian population. In contrast, the more highly conserved genes, *mero2* and *TashHN* showed little difference between the two samples. As *mero1* has been shown to be under positive selection, it is logical to consider which sampling strategy performed best in demonstrating selection on this gene.

As indicated by the *mero1* results, 47 unique alleles were identified in the Turkish sample, which equates to an average of almost twelve genotypes present in each blood preparation. Micro- and mini-satellite genotyping had already estimated the multiplicity of infection in each blood sample by counting alleles at the most informative locus. This varied between nine and twelve genotypes and is consistent with the antigen sequencing data. Since 24 clones were generated from each isolate, as many as 50 % of the sequences generated by this method were unique and therefore informative. Encouragingly both sampling strategies showed a similar pattern of nucleotide diversity along the length of the gene. Within the Turkish and Tunisian sequences, the mean value of π was 0.0468 and 0.0396 respectively, while the value calculated over the entire sequence set was 0.0476. The Turkish and global values are very similar, with the addition of the Tunisian sequences to the Turkish sample set contributing little additional polymorphic information. This can also be seen when analysing the number of polymorphic nucleotide sites in each of the two population samples – 95 for Turkey and 74 for Tunisia, with a global value of 99. In this regard, generation of multiple sequences from a limited number of cattle blood samples (i.e. four piroplasm isolates) would appear the most efficient method for detecting allelic diversity. The contemporaneous sampling of young unvaccinated cattle under a similar management regime ensures that a genuinely sympatric population has been sampled and that the allelic sequences may be compared with confidence. However, a potential

drawback of this approach is the introduction of erroneous nucleotides in the PCR stage of the protocol.

5.4.4. PCR and sequencing errors

All sequences generated in this study were obtained at a level of at least '2x' coverage and involved sequencing in both directions. On average between '3x' and '4x' coverage of every nucleotide was obtained because of the requirement of internal priming sites at around 500 bp intervals. The sequencing reads obtained from the commercial service were an average of around 800 bases of high quality sequence. Therefore, in this study the impact of sequencing errors can be considered to be negligible, due to the sequence of each nucleotide being determined as a consensus from multiple reads. However, this highly accurate sequence data may be derived from clones containing errors introduced by the PCR reaction. The inherent problem with the polymerase chain reaction is that amplified DNA fragments are liable to nucleotide substitution during the reaction process. The rate of these artificially introduced polymorphisms has been previously documented (Lundberg *et al.* 1991) and a recent study has highlighted their influence on molecular population genetics (Kobayashi *et al.* 1999). *Taq* polymerase lacks 3' to 5' exonuclease activity, which renders it unable to recognise and remove mismatched bases as they are added to the growing oligonucleotide. This results in an error rate of approximately 1 in 10,000 bases, which if it occurs early in the amplification reaction can alter large proportions of the final product. Other polymerases with 3' to 5' exonuclease activity, such as the *Pfu* used in this study are available to increase accuracy, however failure to catalyse reactions involving larger DNA targets have been shown with this polymerase (Lundberg *et al.* 1991). However the principal reason for not using pure *Pfu* was in this study was that an un-templated A-overhang was required for the cloning reaction, which the *Pfu* polymerase would have removed. A small amount of *Pfu* mixed with *Taq* was used in this study as it has been shown to improve the efficiency and fidelity of PCR reactions (Barnes 1994).

The ten Tunisian parasite preparations containing a single genotype were used to estimate PCR error rate since the amplification reaction should in theory generate a single species of PCR product. This was found to be the case and allowed a PCR error rate of 0.94 errors per kb to be computed. It is important to consider how errors introduced through PCR can influence the results of the various molecular population genetic tests used in this study. Such errors may influence Tajima's *D* test, as indicated by a study that estimated several of the population parameters commonly used for describing DNA polymorphism across 16 individuals of *E. virgintioctomaculata*, comparing an accurate dataset with one subject to

PCR error (Kobayashi *et al.* 1999). The number of haplotypes, the number of segregating sites (S) and nucleotide diversity (π) were found to be larger from the data representing the PCR-error prone DNA, leading to an overestimate of the DNA polymorphism within the population. Tajima's *D* value was estimated to be 0.51 from the data including the PCR errors, although the correct value is known to be 1.37. This is attributed to large increase in the value of S compared to a very limited rise in π resulting in lower than expected value for the test. DNA sequence errors were found to be randomly distributed through the gene and can be regarded as effectively neutral. If sufficient PCR errors are present in a set of sequences then the effects of positive selection may be masked, and the gene would appear to have evolved neutrally. Logically, this would be predicted to confound tests that are seeking to identify a departure from neutrality, biasing them towards more conservative results.

In the case of *mero1*, a method was used to correct for PCR errors before the sequences were subjected to Tajima's and Fu and Li's tests of neutrality (Table 5.12.). Similar to the results from the Kobayashi study, the number of segregating sites is significantly higher in the error-prone results, 122 compared to 99 in the 'corrected' dataset. Nucleotide diversity remains almost unchanged at 0.048 in the corrected sample compared to 0.049 in the other. Although not all statistically significant, the values of Tajima's *D* and Fu and Li's *D* and *F* indices are increased over the entire dataset and also within the Turkish population. However, when one considers the algorithms used in these tests, this is perhaps not surprising. In effect, Fu and Li's *D* and *F* indices, identify singleton polymorphisms across the dataset and then compare them with their expected frequency based on a neutrally evolving population. The method of error correction used in this study, by definition, locates and eliminates them from the analysis. It could be argued that this artificially excludes evidence of neutrality. Also, by definition, codons that contain polymorphic information are not subject to 'correcting' and may therefore still contain erroneous sites, which are not eliminated. This is a lesser problem, as the polymorphic information for such a codon is likely to outweigh a single chance nucleotide substitution: i.e. the more nucleotide diversity, the lesser the deleterious effect of the PCR error. It is interesting, therefore, to analyse the distribution of the three neutrality indices over the length of the gene. From Figure 5.10. it can be seen that several areas are identified where the index values approximate to zero or drop as low as -1.5 (Tajima's *D*) and –2 (Fu and Li's *D* and *F*). Troughs in value can be seen between bases 0 and 100 and also between bases 500 and 600 and when nucleotide diversity over these areas is examined, it is found to be low (in the first area) and falling from a high value (in the second area). On examination of the

$d_N d_S$ plot (Figure 5.9.(i)) both these areas can be seen to be under the influence of purifying selection. The first coincides with the location of the signal peptide and the second is a short conserved region just before the GPI-anchor motif. The nucleotide diversity is generally low, therefore there should be little masking of PCR-errors, which should be readily identified. To summarise, the tests are still able to achieve low or negative values in the face of the PCR-error correcting methodology.

This method was also used to assess the impact of PCR error on both the McDonald-Kreitman and $d_N d_S$ tests. Corrected and un-corrected datasets for both merozoite antigens and SVSP genes were subjected to each test and the results compared. The corrected and un-corrected data agreed very closely with each other, and no significant differences between datasets were found for either of the tests (results not shown). The limited amount of nucleotide diversity exhibited by the host nuclear genes precluded them from this analysis. Consequently, the results presented in this thesis for the McDonald-Kreitman and $d_N d_S$ tests relate to the original uncorrected sequence data for all the genes except *mero1*.

5.4.5. Merozoite candidate antigens

The merozoite candidate antigens show highly contrasting results using the McDonald-Kreitman and $d_N d_S$ analysis, however, for each gene the results of the tests are consistent with each other. *mero2* seems to be heavily under the influence of purifying selection. Although a moderate level of polymorphism is observed at the DNA level, the amino-acid sequence of the protein is highly conserved. This is most easily seen when examining the $d_N d_S$ plot (Figure 5.9.(ii)), where the vast majority of segregating codons are subject to negative selection. A few positively selected sites are scattered through the gene, however none have statistical significance at $p < 0.1$. Conservation within *T. annulata* is underlined with the McDonald-Kreitman results (Table 5.10.), where intra-species conservation is highly statistically significant raising the question - should this gene still be considered as an antigen candidate? Naturally, arguments can be made for and against. Only a small number of alleles were effectively sampled from each population, which is attributed to in-efficiency in the cloning protocol, as PCR products were successfully generated. However, the fact that PCR products of the correct size were obtained from each template DNA and that amino-acid composition is highly conserved suggests that there is a selective pressure for an invariant protein to be maintained in genotypically diverse strains of *T. annulata*. The GPI-anchor and signal peptide motifs are conserved, and if these bioinformatic predictions, along with the EST data, are accurate then the protein is likely to ubiquitous (among strains) on the merozoite surface. This would predict exposure to the

bovine immune response and selective pressure to diversify. When one considers that cattle can become immune to tropical theileriosis, it is necessary to consider that this implies the parasite is ineffective at evading the bovine immune response. One could speculate that there may be a single or set of highly conserved proteins that confer the type of cross-protection elicited by natural infection. This is inherently unlikely, and the explanation for its lack of allelic diversity, and lack of evidence for immune selection, is likely to be much more mundane. Explanations for an apparent lack of diversification include (a) the GPI-anchor prediction algorithm may be inaccurate, in which case the protein would be localised inside the cell and never exposed directly to the host immune system and (b) possession of a tight globular conformation, whereby mero2 is shielded by the 'fuzzy coat' of the merozoite, provided by a layer of TaMS1 or other merozoite surface proteins (Katzer *et al.* 2002). To investigate this, it would be necessary to determine its sub-cellular localisation and this would involve expressing recombinant protein and generating anti-sera. Another factor to consider is the robustness of the EST data. In this study, the available EST database was only represented by qualitative data, i.e. presence or absence of hits with respect to one of the three life-cycle stages. In theory, a single, spurious EST hit would be enough to indicate merozoite expression, although expression may be transient or at a relatively low level. Additionally, it is notable that the *mero2* EST corresponds to expression in the merozoite stage alone, in contrast with *TAMS1* and *mero1*, which are expressed in both the merozoite and piroplasm stages. It is possible that diversity in *TAMS1* and *mero1* is the result of exposure to the tick environment in the piroplasm stage. It is difficult to believe that the arthropod immune system itself could contribute diversity within *T. annulata* for two reasons – (1) the tick is likely to be incapable of mounting specific responses to variant parasite antigens and (2) each tick participates in a single generation of the life-cycle of *T. annulata* genotypes and acquired immunity (if actually possible) would not be present when nymphs are infected during their blood meal. If bovine immune selection actually occurs in the tick (through co-ingestion of piroplasms and antibodies), then only those proteins expressed on the piroplasm surface will be subjected to a pressure to diversify. These hypotheses will hopefully be investigated in the near future when the quality of the EST data can assessed with the advent of a *T. annulata* micro-array study.

In contrast to *mero2*, *mero1* appears to be an antigen and may be a promising vaccine candidate for several reasons -

1. Presence in all strains. In common with *mero2*, the primers amplifying the *mero1* gene were able to generate PCR product using each DNA templates in the study. This suggests that the protein is likely to be present in diverse parasite strains and therefore may be an effective target across different regions of the world.

2. Limited structural polymorphism. Only two size variants of this gene have been identified across almost 60 allelic sequences, which differ by only 3 bp, encoding aspartic acid at codon number 170. This level of length conservation is second only to *mero2* in the six genes in the study. Both the signal peptide and GPI-anchor regions show strong evidence of purifying selection, suggesting these regions are critical for the effective functioning of the protein (Figure 5.9.(i)). The orthologous gene in *T. parva* shows very high identity with *mero1* at the nucleotide and amino acid level – 85 % and 83 % respectively, suggesting that there may be selection acting to conserve this product of this gene across species. This high level of conservation suggests that it may have an essential function for the survival and transmission of the parasite. A possible function could be in the recognition and invasion of the host erythrocyte such that mutations leading to loss of function would result in failure to complete the life-cycle.

3. Evidence of merozoite surface location. The orthologue of mero1 has been demonstrated on the surface of *T. sergenti* piroplasms (Zhuang *et al.* 1995) using a monoclonal antibody, which recognises the 23 kDa orthologous protein. Two-dimensional polyacrylamide gel electrophoresis also confirmed that this protein and the TaMS1 orthologue were expressed in all *T. sergenti* isolates examined. This evidence of surface location in other *Theileria* species suggests that the GPI-prediction and EST data are correct and that mero1 is genuinely expressed on the merozoite and piroplasm surface in *T. annulata*. This could be confirmed experimentally in *T. annulata* by generating antisera to recombinant mero1 protein, which could then be used to detect the native protein in merozoite preparations by Western blotting and IFAT.

4. Evidence of positive selection. The McDonald-Kreitman and $d_N d_S$ results indicate an excess of non-synonymous substitutions within *T. annulata*. The $d_N d_S$ value is much higher than that calculated between C9 and *T. parva* (see Table 5.11.). Five sites under positive selection ($p < 0.25$) were identified, several of which cluster around the inserted / deleted codon (Figure 5.9.(i)). However, consistent with a general background effect of conservation, a large number of synonymous substitutions are also evident. The gene displays a very similar pattern of diversity to that in *TaMS1*, although the absolute number of positive sites is around three times higher for *TaMS1*. However, the results of the

McDonald-Kreitman tests for *mero1* are more conclusive. More non-synonymous than synonymous are found for both genes within *T. annulata*, however the ratio is inverted when comparing *mero1* between the species. This explains the higher value for the neutrality index and a much lower *p* value for Fisher's exact test. In other words, there is compelling evidence that the composition of *mero1* has been constrained between the species, but that it is being driven to diverge within *T. annulata*. This is consistent with and good evidence for immune selection. Interestingly, this trend is less evident when the McDonald-Kreitman test is performed using the other *Theileria* species. A significant amount of sequence was discarded in order to generate an un-gapped multiple alignment with these species. Lower protein identity (< 60 %) and that fact that these species are more phylogenetically distant than *T. parva* may reduce the power of the test to detect departures from neutrality. It is particularly notable that *mero1* outperforms *TaMS1* in the search for positive selection. The high level of sequence polymorphism and non-synonymous nucleotide differences has been proposed as evidence that *TaMS1* would be a good vaccine candidate (Gubbels *et al.* 2000b). However, the evidence that diversifying selection is operating on this molecule poses a problem in terms of its use as a vaccine antigen as any induced immune response will need to recognise a large number of diverse alleles. As previously mentioned, the high amino-acid identity between orthologues of *mero1* (83 %) is more indicative of a constraint on divergence than the value of 77 % for *TaMS1* and therefore it might be less able to evade a protective immune response. For further evidence of such a constraint, the influence of balancing selection needs to be reviewed.

5. Evidence of balancing selection. After PCR error correction, 99 segregating sites were found in the DNA sequences generated for *mero1* and these are shown in Figure 5.7. This represents the polymorphic data present among all the alleles, which is the underlying variation that is used in the tests of neutrality. All three neutrality indices returned positive values for *mero1*, in contrast to *TaMS1*, which displayed negative values, however PCR errors are likely to be present in the latter dataset, which would likely nullify any balancing selection as previously described. For *mero1*, Tajima's *D* and Fu and Li's *F* test return values above the upper limit of the 95 % confidence intervals based on free recombination between genotypes, therefore the null hypothesis of neutrality can be rejected. The free recombination confidence interval is narrower than the confidence intervals using the more conservative 0 % recombination rate. However, the free recombination model is suitable for this study for two reasons. Firstly, *T. annulata* is understood to be panmictic (Sections 2.4.2. and 3.4.1.), with an obligate sexual cycle taking place in order to produce

each new generation. Secondly, as population sampling is conducted at such fine resolution, with neighbouring cattle being infected by a single population of ticks, it is reasonable to assume that a genuine mixis of genotypes is taking place. This is supported by the dendrogram depicted in Figure 5.8.(i), which suggests a lack of any sub-structuring within the Turkish population, with the major clusters incorporating sequences from all four cattle. The results of Tajima's and Fu and Li's tests therefore indicate that *mero1* is under the influence of balancing selection and that a limited number of genotypes are maintained within the population with novel mutant genotypes not arising as would be predicted if the gene were neutral. This balancing selection is consistent with frequency-dependent selection acting on the antigen from the bovine immune system.

Taken together, these factors provide strong circumstantial evidence that *mero1* is an antigen and that its allelic diversity is consistent with bovine immune selection. The potential utility of the molecule as a component in a sub-unit vaccine will be discussed more broadly in Chapter Six.

A degree of caution should be exercised when analysing the results of these tests following the PCR-error correction process. By removing point mutations from the sequence data, one may actually be obliterating genuine mutations present in the parasite genome. Whereas the original uncorrected data may be biased towards neutrality, PCR-error corrected sequence may be biased towards a departure from neutrality and balancing selection. However, the error-correction protocol used in this study is relatively conservative and it is anticipated that only a limited number of genuine neutral mutations were discarded. To be truly confident that PCR errors are absent from the sequences, an alternative experimental protocol is necessary. One method would be to perform multiple PCR reactions from each template and generate clones for sequencing for each of them. The PCR products are generated from independent reactions and therefore it is highly unlikely the same error will be reproduced. Three PCR reactions, would allow a consensus of two identical nucleotides to be achieved at each base position. This would have been possible using the Tunisian samples in the present study, however the heterogeneous nature of the Turkish samples would have necessitated a great deal of time and expense in generating and sequencing a very large number of colonies.

5.4.6. Diversity and selection in two macroschizont gene families

The findings presented in Chapter Three of this thesis represented a novel bioinformatic method of identifying positively selected genes, by comparing the genomic sequence data

of *T. annulata* with that of *T. parva*. As predicted from previous studies, high interspecies non-synonymous to synonymous substitution rates were observed in known antigen genes, however this was also associated with two classes of gene not previously identified as encoding antigens. These were (a) a family of parasite-encoded host nuclear proteins (TashATs) and (b) the sub-telomeric variable secreted protein (SVSP) family. These *Theileria*-specific protein families show a degree of relatedness to each other and are considered to be subsets of a larger super-family of genes. The host nuclear genes *TashHN* and *SuAT1* were included in the allelic sequencing study primarily as a form of control, on the assumption they would not be under diversifying selection. It was expected they would be highly conserved within *T. annulata* since they are believed to undertake specific, critical functions in modulating host cell phenotype (Swan *et al.* 2001; Swan *et al.* 2003; Shiels *et al.* 2004). Accordingly, it was anticipated that their diversity would contrast with that of putatively polymorphic merozoite antigen candidates.

TashHN was found to have very limited allelic diversity both within and between countries, as presented in Table 5.9. This low level of polymorphism (only 20 polymorphic nucleotides) reduced the power of $d_N d_S$ analysis to determine whether codons were genuinely under positive or negative selection. Despite this, five codons were found to be under the influence of negative selection at $p < 0.25$, whereas none showed evidence of positive selection, raising the question - how should lack of polymorphism in *TashHN* be interpreted? It could be argued that all the invariant nucleotides were highly conserved with any mutations, even those resulting in silent amino acid substitutions, being selected against. This would require a strong mutational bias in this gene to revert substitutions to their original state. However, this is unlikely, since *TashHN* would need to be composed almost exclusively of a set of preferred codons with mutation to non-preferred codons, being highly detrimental to the parasite. A more likely explanation is that the gene has recently evolved. As previously discussed, *mero2* is highly conserved at the amino acid level, which is associated with a high level of synonymous substitutions at the DNA level between alleles. Why isn't a similar level of synonymous substitutions observed in *TashHN*? If the genome is subject to a constant rate of neutral mutation then one would expect a similar level of synonymous substitutions in both genes. This raises the notion that perhaps both genes did not evolve at the same point in time. It is beyond the scope of this project to investigate this temporal origin of these loci, however a few observations can be made. When the *T. parva* orthologues of these genes are compared, a very similar level of nucleotide identity is displayed – 77 % (see Table 5.1.). Protein identity is lower in TashHN at 65 % compared to 71 % in mero2. Both the ancestral genes must have been

present in the common ancestor of modern day *T. annulata* and *T. parva* and it is possible that the variable level of polymorphism displayed by each these two loci is evidence of multiple events when buffalo derived *T. parva* 'jumped' into cattle. To explain, if a single haploid parasite had established in cattle then this would represent a single founder genotype. Consequently, all diversity in current populations of *T. annulata* would have arisen since this speciation event. However, if multiple genotypes had established in cattle then pre-existing polymorphism in the buffalo parasite would be present in the *T. annulata* founder population. A similar level of DNA polymorphism between the orthologues of both genes suggests a comparable mutation rate being in effect since *T. annulata* and *T. parva* separated. One may then predict a significant level of polymorphism pre-existing in the *mero2* sequence when the 'jump' took place. However, *TashHN* may have evolved just before the species diverged and therefore may not have been polymorphic at this point in time. With *TashHN* presumed to be an ancestral gene in the TashAT locus, this would imply that the TashAT family as a whole are relatively recent in evolutionary terms. The evolution and function of this gene family raises questions such as – is this family present in more distant species like *T. sergenti*, which are unable to transform host cells and does *mero2* have an orthologue in such a species? To further investigate the evolution of this locus, it would be informative to determine the allelic variation across the TashAT family within a panel of *T. parva* isolates. However, to summarise, the present results suggest that *TashHN* and the TashAT locus may still be evolving, while the function of *mero2* is fixed in *T. annulata*.

Similar to *TashHN*, a low level of diversity was also encountered for $SuAT_1$ when each country was analysed separately, however, significant divergence was observed between the two countries. In addition to Tunisian-type and Turkish-type sequences, intermediate-type sequences were identified. In a sense, the variation of these two genes reflects that exhibited by particular micro- and mini-satellite loci. *TashHN* is analogous to TS15, where a small number of alleles are identified overall and there is limited evidence of differentiation between countries, whereas $SuAT_1$ is more akin to TS25, where particular alleles clearly predominate in each country (Figure 3.5.). The variant behaviour of the two host nuclear genes highlights the need to study multiple alleles of several genes when attempting to draw conclusions about the nature of diversity in different classes of gene. Diversity at the TashAT locus exhibits high inter-species $d_N d_S$, but within *T. annulata* the diversity of the genes studied has clearly been constrained. This is particularly evident across the putative functional domains, which account for nuclear localisation properties and where binding to host DNA is predicted to occur. Additionally, the functionally

uncharacterised motif, L(Q/V)PETIPVE, was shown to be strongly conserved in $SuAT_1$. It could be argued that because they encode several closely related proteins, the host nuclear family possesses a degree of redundancy. However, this study has shown that both *TashHN* and *SuAT1* are present in every isolate of *T. annulata* and are broadly invariant. This suggests that not only are they critical to the biology of the parasite, they are highly specific in their function with novel mutations apparently being quickly eliminated from the population. Despite this observation, the study has identified a divergence between two principal allelic types of *SuAT1*. It is possible that the allelic types have evolved in different environments and adaptation to different tick species has been proposed. However, the simplest explanation is that divergence is due to a degree of isolation and genetic drift, whereby mutations in the gene arising in separate areas eventually resulted in two allelic types that were unable to recombine to produce a viable hybrid allele. This finding is consistent with the population genetics studies where a moderate amount of genetic differentiation was demonstrated between populations in Tunisia and Turkey (Section 3.3.2.). Neither of the host nuclear genes sequenced in this study showed evidence of positive diversifying selection in order to evade the bovine immune system, with the high intra-species level of $d_N d_S$ of *SuAT1* attributed to comparing alleles of different type. Therefore, it can be concluded that the high inter-species $d_N d_S$ values of host nuclear genes are due to 'directional' selection, with the locus continuing to evolve independently in *T. annulata* and *T. parva*. This stands in contrast to the pattern of diversity shown by the SVSP genes.

The SVSPs are a large sub-telomeric gene family having signal polypeptides and PEST domains and are located on all telomeres. To date, no information is available about their localisation within the parasite or infected host cell and their function is an enigma. Similar to the TashAT genes, the family was identified as possessing high inter-species $d_N d_S$ values. To investigate the significance of this observation, two representative members were studied in detail with alleles sequenced from both Tunisian and Turkish populations. These selected genes were typical of the family, possessing a signal peptide, multiple PEST motifs and a high inter-species $d_N d_S$. A great deal of diversity was encountered at the nucleotide and amino acid level across the length of both genes, with hypervariability identified around the PEST motifs. This is in sharp contrast to the conserved host nuclear genes and clearly indicates the SVSP genes are not subject to the same constraints. Similar to the host nuclear genes, convincing orthologues were identified for around half of the family members, indicating that a high proportion of genes have been maintained in both parasite species. This is perhaps surprising given (a) the

allelic diversity revealed in this study and (b) the fact these genes appear to be evolving in a genuinely neutral manner. Although there is generally more non-synonymous than synonymous substitutions, they effectively arise at the same rate when account is taken of the number of potentially non-synonymous and synonymous sites in the molecule. Indeed, the McDonald-Kreitman and $d_N d_S$ analyses indicate that although highly variable, they are probably not under the influence of a strong positive selective pressure and should perhaps instead be considered as being unencumbered by purifying selection. This is not purely a difference in semantics. It is generally accepted that although genes mutate and evolve through largely neutral processes (Kimura 1979), there must be a background level of conservation that constrains proteins from constant radical alterations affecting their function. This can be most easily observed at the species level where the underlying effect of purifying selection acts across the genome. Unusually, the SVSP proteins appear relatively free of this restraint with a large proportion of neutral mutations giving rise to viable alleles. Why, then should SVSPs be neutral, or perhaps more pertinently, why is this background level of purifying selection undetectable? To date, there is no evidence that the function of SVSP proteins is dependent on folding activity and tertiary structure. The data suggest that there may be little constraint on the distribution or quality of the amino acid replacements, even when this may result in a radical change to the three dimensional conformation of the molecule. Moreover, the multi-copy nature of SVSPs may be indicative of a degree of redundancy across the family. One may speculate that the underlying factors promoting a multiplicity of SVSP loci and diversity between these loci also promote allelic diversity between genotypes. Pseudogenes may be one of the few classes of sequence that may be considered to be evolving neutrally. Since pseudogenes do not encode expressed proteins, there would be no constraint on their composition because they do not perform any function. However, totally neutral mutations would be predicted to result in random stop codons and frame shifts in the DNA sequence. Since the SVSP family all possess open reading frames, this is unlikely to be the case. Whatever their function, the multiplicity, polymorphism and expression profiles of SVSP genes suggest they are essential for the *T. annulata* macroschizont.

It was suggested in Section 4.4.5. that SVSP proteins may be involved in evading the host immune response. Moreover, the location and arrangement of SVSP genes were discussed in relation to contingency genes in other protozoan species in Section 5.1.4.1. Such parasite genes are often associated with the telomeres and it is theorised that sub-telomeric location provides the capacity for gene family diversification through ectopic recombination (Barry *et al.* 2003). Although the chromosomal location of the SVSPs may

provide a mechanism for generating diversity at a higher rate than interstitial genes, it does not explain the underlying reason for this variation. It was suggested in Section 4.4.5 that since SVSP proteins contain a signal peptide and multiple PEST motifs, they may be secreted in to the host cell compartment, degraded and presented to the bovine immune system on MHC Class I molecules. However, the fact that *SVSP1* and *SVSP2* appear to be evolving neutrally suggests that the bovine immune system is not driving them to diversify. A simple hypothesis may explain this apparent paradox. SVSP genes are multi-copy and highly variable and this coupled with a high level of allelic diversity means that the parasite may encode a vast array of different peptides within and between genotypes. This would mean that any one particular locus would not be under strong selective pressure in the population. However, it is important to consider why it may be necessary for the parasite to secrete proteins for presentation on the host cell, when the macroschizont is an intra-cellular stage, effectively shielded from the immune response.

It is possible that SVSPs may function to mask proteins, which are secreted from the macroschizont and are thus at a high risk of cleavage and presentation by the host cell. As previously discussed, the TashAT family of proteins are secreted and travel to the host cell nucleus in order to exert control over the leucocyte. TashAT and SVSP proteins share a degree of homology and it is likely they belong to larger family, which comprises subsets of proteins that have diverged to perform specialised functions. This family also includes the major macroschizont surface protein, TaSP. However, the fact that TaSP is tethered to the macroschizont surface suggests that it may not be presented on the surface of the leucocyte in high quantities. For peptides representing this molecule to be presented, the protein must either be degraded *in-situ* on the surface of the macroschizont or detached before being degraded while free in the host cell cytoplasm. TashAT proteins, like SVSPs, possess PEST sequences that signal them for proteolysis. *TashHN* and $SuAT_1$ were shown to be under the influence of purifying selection, therefore one may predict that they would present a limited set of peptide fragments, which can pass into the MHC Class I processing pathway. Consequently, these relatively invariant peptides may be presented and recognised by a protective immune response. However, it may be hypothesised that SVSP peptides act as competitive inhibitors, blocking presentation of protective epitopes. This may reduce the concentration of TashAT fragments presented to the immune system and may preclude a protective response targeting the infected leucocyte. Disappointingly, recent vaccine trials based on recombinant TashAT and TaSP protein did not stimulate protective immunity (unpublished data). Nevertheless, BLASTP analysis of the *T. annulata* genome has shown that SVSPs have greater similarity to many hypothetical

proteins than to TaSP or TashAT family members. This suggests that there may be a completely novel set of proteins that have the potential to be masked by SVSP peptides. However, since many of these proteins are genus or species-specific, they have yet to be characterised and their function remains unknown.

Further studies require to be undertaken to investigate the relationship of the SVSP family with proteins expressed in the macroschizont, to either support or refute the masking hypothesis and to highlight the protective antigenic domains. If this hypothesis is valid, it is unlikely that the SVSP would be of any value as a component of a sub-unit vaccine, since limited cross-protection would be generated by stimulating the bovine immune system with a single SVSP allele. Although the immune system may recognise SVSP alleles individually, an animal vaccinated with such a protein would only eliminate parasites harboured in leucocytes containing macroschizonts expressing an identical or closely related SVSP allele. If extensive cross-protection was detected, then the masking hypothesis would be cast into considerable doubt. Additionally, the *T. annulata* genome possesses a large family of SVSP paralogues, and it is possible that a proportion of the parasite population in an infection may express alternate SVSPs to the immunising gene. However, current EST data, based on a single macroschizont preparation, suggests that the majority of SVSPs may be expressed simultaneously. There is a clear need to investigate the expression characteristics of the SVSP family in detail. With the advent of a *T. annulata* micro-array study in the near future, hopefully two basic questions can be addressed - (1) is there variation in SVSP expression in relation to parasite genotype and (2) are there temporal differences in SVSP gene expression over the course of infection? This may reveal either a cyclic pattern of protein expression, where different loci are expressed in turn or alternatively it may indicate that the parasite takes a less subtle approach to immune evasion, where the majority of macroschizonts are expressing the majority of genes. Additionally, it would be useful to experimentally investigate whether SVSP peptide fragments are presented on MHC Class I molecules. Elucidating SVSP expression and localisation is therefore critical to understanding if this gene family does indeed play a role in immune evasion.

CHAPTER SIX

GENERAL DISCUSSION

6.1. Importance of studying field populations

Over the course of the last century, tropical theileriosis has emerged as a major constraint on livestock production in tropical and sub-tropical regions of Europe, Africa and Asia. Following identification of the aetiological agent at the beginning of the 20th century, work was primarily directed toward investigating the basic ecology and biology of the parasite. However, as the full impact *T. annulata* infection on cattle health was appreciated, research focussed on developing prophylactic measures in laboratories in endemic regions. In the 1960s, this culminated in the development of cell line vaccination based on the finding that the parasite could be maintained indefinitely through *in vitro* culturing (Tsur and Pipano 1966). This advance has afforded researchers in non-endemic regions the ability to study the parasite in great depth. By necessity, much of this work has involved studying genotypic and phenotypic traits of a limited number of parasite isolates or clones. The capacity for *T. annulata* to be maintained *in vitro* has been a double-edged sword, in that it has allowed a single collection of parasite material to be re-visited for separate studies and has obviated much of the need to collect new isolates. For example, the considerable amount of parasite material in the form of cell lines and DNA collected in Tunisia in 1993 has been used as a resource for several studies on the diversity of *T. annulata* (Ben Miled *et al.* 1994; Katzer *et al.* 1998; Taylor *et al.* 2003). Although one may have expected the data on each isolate to accumulate across each successive study, the two latter studies represent independent analyses without integration of previous data. This particular collection of samples was obtained in a structured manner representing isolates from the within the same site and across different sites in the North of the Tunisia and remains an indispensable resource for analysing variation within this region. However, with the wide distribution across the world of endemic countries, absolute conclusions regarding parasite diversity in the wider context must be viewed as speculative when drawn from the Tunisian dataset alone. For instance, the study on allelic sequences of the *TaMS1* gene (Gubbels *et al.* 2000b) analysed these same Tunisian isolates along with a disparate collection of isolates from countries throughout the world. Although this study was highly informative, charting extensive sequence variability at this locus, it concluded that there was no clustering of sequence types with their place of origin. While this conclusion was

valid with respect to those isolates used in the study, no general inference could be drawn relating sequence diversity to geographical origin using that particular dataset.

In contrast with previous studies, the results presented in this thesis represent the first analysis of *T. annulata* diversity using an extensive dataset from multiple geographic areas. Paradoxically, it was the advent of the genome-sequencing project, which stimulated this study. Analysis of the genome sequence of the single clone of *T. annulata* allowed the identification of the micro and mini-satellite loci, which were likely to be highly polymorphic and genetically informative. The identification of these loci facilitated the relatively rapid development of molecular markers, which were applied to address questions regarding the basic population biology of the parasite, which had previously remained unanswered. The collection of Tunisian isolates mentioned above was analysed to demonstrate those loci that were polymorphic and was used in the preliminary population genetic analysis of isolates from Tunisian and Turkey. In part, the results generated in this study were related back to the original genotyping studies undertaken by Ben Miled (Ben Miled 1993; Ben Miled *et al.* 1994). This was done principally for two reasons – (1) it allowed a more comprehensive understanding of the data and (2) it served as a form of internal quality control, whereby the results could be reconciled with previous findings.

Since the genotyping system is PCR-based, small quantities of parasite DNA may be used as a template for amplification. The development of the related technique of human genetic fingerprinting in the 1980s allowed law enforcement authorities around the world to examine archive biological material for DNA analysis (Jeffreys *et al.* 1985b; Jeffreys 2005). In many instances, this has proved to be very successful in matching query material against a database of genetic profiles. In a sense, this is analogous to the application of the *T. annulata* micro and mini-satellites to the field populations of the parasite in this study. In both Tunisia and Turkey a large archive of blood samples from infected and carrier animals has been collected and stored over the last decade. The development of the new markers has allowed this material to be re-visited and the parasite genotypes analysed. This has lead to five inter-related conclusions about the basic diversity and population genetics of *T. annulata* -

1. There is an association between parasite origin and genotype. Whereas no association had previously been established between the geographical origin of parasites and genotype, cluster analysis with the new markers discriminated between isolates from Tunisia and Turkey. This finding was indicated in the preliminary analysis when a limited

dataset was available, though it was necessary to generate and compare full genetic profiles representing each parasite mixture before this could be achieved. However, when the two more extensive field populations were analysed, PCA clearly demonstrated an association amongst isolates from the same country (Figure 3.8.).

2. Genetic differentiation is detectable between geographically distant regions. A moderate amount of genetic differentiation was confirmed between the substantial collections of *T. annulata* isolates gathered in Tunisia and Turkey, which was reduced when sampling sites within each country were compared (Table 3.4.). Inter-country F_{ST} values calculated in the preliminary study agreed closely with this analysis. Quantification of the differentiation with this standard population genetic statistic is consistent with the observed association between parasite genotype and geographical origin.

3. Isolates from within a locality show a degree of relatedness. When similarity analysis was performed on isolates collected from each of the four districts in Western Turkey, the highest identity between pairs of samples was found between isolates from within the same district (Table 3.7.). Although, these districts were not genetically isolated and did not group independently (Figure 3.12.), an element of clustering of parasite genotypes from the same district was observed. In other words, there is an association between parasite origin and genotype across limited geographical distances.

4. Each parasite isolate represents a diverse mixture of genotypes. A high multiplicity of parasite genotypes was identified in every single field isolate analysed in this study (Table 3.2.). Between locations the mean number of alleles per locus varied between two and four (Figure 3.16.(i)), though this provided only an index of the actual heterogeneity in each isolate. An alternative parameter, the maximum number of alleles at any single locus indicated that the mean value, as used in this study, is likely to be a highly conservative estimate. Indeed, a cut-off value of twelve alleles had to be imposed when abstracting allelic data from several highly diverse isolates.

5. The number of harboured genotypes is related to the age of the host. A statistically significant positive correlation between host age and the multiplicity of infection was identified in four out of the five locations analysed (Figure 3.15.). Isolates collected at El Hessiène, Tunisia demonstrated this within the first disease season, while three districts in Western Turkey indicated this trend across several seasons of exposure. These results are consistent with the view that non-sterile immunity of cattle to tropical theileriosis in endemic areas is maintained by constant re-challenge (Ilhan *et al.* 1998).

This study underlines the need for an appropriate sample collection, designed to address a particular set of questions. Several factors contributed to the study being able to arrive at these conclusions – an extensive number of isolates were collected; the collection represented two structured sampling programmes, whereby a considerable number of isolates were collected from within discrete localities; a suitable distribution of cattle (with respect to age) was sampled; and most importantly the blood samples had been suitably stored and catalogued, with an extensive amount of pertinent host data recorded at the time of collection. In addition to the marker-based study, a sequence-based approach was undertaken to determine diversity at six loci within the *T. annulata* genome, purported to be under the influence of positive selection. The multiplicity of infection indicated by the polymorphic neutral markers had suggested a single mixed blood preparation would represent a large number of allelic sequences of these genes. Consequently, micro- and mini-satellite genotyping was utilised to identify a number of highly heterogeneous isolates from the village of Sariköy in Akçaova, Western Turkey. The primary goal of this sequencing project was to determine whether positive selection, suggested at the inter-specific level, was reflected intra-specifically. Thus, overall, the work presented in this thesis represents two alternate approaches of characterising diversity in the genome of *T. annulata* and their application at different levels of resolution to address different questions. In the case of the marker-based study, basic population genetic questions were addressed by analysing diversity on an extensive geographical scale. In contrast, a small number of isolates (four) from one locality were used to generate dozens of sequences for comparative allelic analysis. Selection, especially immune selection, is a local phenomenon and therefore, the identification of these four heterogeneous samples was critical. Originating from a single sampling site, they would represent parasite genotypes that had been subjected to putatively identical selective pressures. Additionally, a slightly different approach was used to generate some of the allelic sequence data documented in Chapter Five. This relied on the allelic sequencing of several clonal isolates of *T. annulata* and again the neutral micro-satellite markers proved important for the selection of isolates from the panel of available isolates.

6.2. Development and application of genotyping system

A critical factor for the creation of the micro- and mini-satellite genotyping system was the availability of the genome sequence of the C9 strain of *T. annulata*. Although development of the markers started around two years before the publication of the fully annotated genome, large contigs of the sequence data were available for bioinformatic

screening. This allowed the rapid identification of micro- and mini-satellite sequences and so avoided the isolation of satellite markers that, in the absence of genomic sequence, would have been a laborious and time-consuming exercise, involving screening genomic libraries with probes representing short tandem repeats. Such a screen would have been likely to yield only a limited number of loci, comprising di- and tri-nucleotide repeats and consequently only markers TS9 and TS16 would have been identified. The availability of a diverse panel of isolates was also crucial to identifying markers that could consistently amplify from parasite DNA and provide information on polymorphism. The identification of these micro- and mini-satellite markers allowed the investigation of the population genetics of *T. annulata* and follows the success of the same approach with the closely related *T. parva* (Oura *et al.* 2003; Oura *et al.* 2004; Oura *et al.* 2005). In both *Theileria* species, markers were distributed across all four chromosomes and a greater number of mini-satellites than micro-satellites were identified. One hundred and thirty-three loci were located in the genome of the Muguga strain of *T. parva* (Oura *et al.* 2003), significantly higher that the thirty-three loci identified in this study. However, the criteria for defining these loci in the *T. parva* were slightly different from those used in this study, therefore a direct comparison between the results is impossible. Additionally, the *T. parva* study was based on the assembled genomic sequence whereas this study was based on the analysis of relatively short contigs. In this study, therefore, some repeat regions may not have been identified because if an entire repeat sequence was not on a single contig, the software may not have detected it. Additionally, in several instances successfully identified repeat loci were discarded because they possessed insufficient flanking sequence to design primers. Sixty of the markers in the *T. parva* study were ultimately shown to be species-specific and polymorphic, contrasting with the ten identified in this study. This can be explained in part by technical differences between the two systems - due to the nature of the size cut-offs used in each analysis. In *T. parva*, alleles could range in size up to at least 600 bp, whereas the upper thresh-hold in this study was 500 bp, which was limited by the Genescan™ size markers. However, it would be interesting to reanalyse the complete annotated genome of both parasites using the tandem repeat finder software but employing identical parameters to identify loci. The actual method for separating and sizing alleles also varied between the studies. The Oura study separated PCR products using 'Spreadex' gel electrophoresis, which is basically an optimised form of the agarose gel method and provides resolution to 3 bp. The use of a capillary-based sequencer in this study allowed accurate sizing to 1 bp, offering several advantages over the gel-based system. This not only allowed the exact size of alleles to be determined, but where multiple alleles were present they could be easily discriminated and enumerated. The

relative abundance of PCR species could also be calculated by the intensity of the fluorescent peak on the trace, and the major product could be consistently identified. In contrast, manually determining the major allele on a gel image introduces an element of subjectivity. Software analysis may assist, but again the resolution of this technique is much poorer than the sequencer-based method. Paradoxically, the high degree of resolution of the Genescan™ traces creates a novel problem. For some markers, alleles are clearly separated by approximately a motif length, whereas other markers exhibit a gradation of PCR product sizes, which means in some regions of the allelic spectrum, there is no clear-cut demarcation between PCR products of differing size. It is therefore necessary to analyse the distribution of PCR product sizes and create suitable 'bins' for allele scoring. In these problematic regions, bins were generally 0.8 bp in range, with PCR products varying in size by as little as 0.2 bp being scored as different alleles. The highly polymorphic nature of many of the markers coupled with this high-resolution sizing method resulted in a plethora of alleles being identified at many loci. Using a less discriminatory gel-based method would have resulted in the scoring of several allelic variants as the same allele, a 'natural' system of binning. So does the analysis system used in this study provide excessive resolution? Logically, the answer to this question must be no. By relying on a gel-based image, polymorphic information is lost through inherent limitations of the technique. This results in a fewer number of alleles being identified but would increase their apparent frequency. However, it would have been equally possible to increase the bin size and to define a lower number of alleles when using the sequencer-derived results in this study. This option was considered, but ultimately rejected for the following reason, if one considers a marker with a complex pattern of diversity, say the 10 bp repeat of TS20 (Table 2.4.), a degree of internal structure is observed, with an ATT motif repeated twice in the 10 bp sequence. If a particular allele mutated on three occasions by the successive addition of a single copy of this tri-nucleotide motif, this would result in a novel allele that was 9 bp larger. If this particular allele then lost an entire 10 bp motif, the resultant allele would differ from the ancestral allele by only a single base pair. Consequently, using an extended bin size, which included products 1 bp apart, would have resulted in the identical scoring of these alleles, which are in fact more distantly related than their intermediate allelic states. The binning technique relies on the supposition that alleles of a similar size are more closely related than ones with a larger size difference. Although this is generally the case, it is entirely possible that over the course of time, identical size-scored alleles may arise through homoplasy. The nature of the genotyping system, whereby alleles are purely defined on their size means this is an

inevitable consequence. Moreover, minor variations seen in allele size may in some cases be due to nucleotide sequence variation. However, without resorting to fully sequencing a large selection of alleles, this particular issue cannot be resolved. A better approach would be perhaps to identify new micro- and mini-satellite loci where alleles are easily scored. By relaxing the criteria for identifying repeat regions, it is highly likely that an increased number of loci would be identified, as was the case in *T. parva* (Oura *et al.* 2003). The initial screening process could then be modified to select loci where either large and / or consistent intervals are identified between alleles, i.e. micro-satellite markers evolving in a step-wise manner. However from this and previous studies (Oura *et al.* 2003), it may be inferred that this type of locus is less frequent in the genomes of both *Theileria* species than mini-satellites, since only four of the ten markers selected in this study were micro-satellites and in *T. parva*, only eleven of sixty markers were micro-satellites. Notwithstanding this, generating new markers would be highly desirable and would facilitate some technical modifications to the genotyping system.

Currently, a separate amplification reaction is required for each marker, and therefore ten separate PCR reactions and sequencer 'runs' are required to generate a full multilocus genotype. Performing multiplex PCR, whereby several sets of primers co-amplify from a single template DNA preparation would considerably reduce the number of PCR reactions involved and decrease the amount of reagents required. This would entail several sets of primers being labelled with different fluorochromes in order that alleles representing several markers could be distinguished on a single electrophoretogram. In addition to decreasing throughput time, cost per sample would be dramatically reduced. However, developing such a system is not a trivial matter and several issues need to be addressed. For example, when an excess of fluorescently-labelled PCR product is analysed, there is 'bleed through' effect, creating artefactual peaks, which would appear to correspond to amplicons labelled with an alternate fluorochrome. Consequently primer combinations must have a similar sensitivity in order to generate PCR products to a similar concentration. The level of polymorphism exhibited by many of the loci described in this study would likely confound such a multiplexing protocol, resulting in traces of immense complexity. Novel, less polymorphic markers would be easier to score as they would exhibit a limited number of well-defined alleles and could be analysed by agarose gel electrophoresis. This would allow implementation of the genotyping system in endemic regions of the developing world where sequence-based assays may not be possible. Additionally, minor size differences in PCR products may resolved by using synthetic polymer technology such as 'Spreadex' gels (Elchrom) (Luqmani *et al.* 1999). In the

course of this project, the markers were used to analyse DNA prepared from frozen blood samples. With the use of FTA filters to preserve blood spots becoming an increasingly effective method of storing parasite material (Rajendram *et al.* 2006), it would be useful if the markers could be applied to this source of DNA. This may involve increasing the sensitivity of the markers and is likely to require the development of a nested PCR protocol. As the majority of field isolates in this study were from clinically affected animals, they would be expected to be relatively rich in parasite DNA and so not present a problem in this respect. However, for some epidemiological applications, increasing the sensitivity would undoubtedly be necessary in order to amplify from cattle with a low parasitaemia.

A large number of cycles of amplification are used in the current PCR regime, in order to maximise sensitivity. However, as the sequencer-based detection of amplicons is in itself highly sensitive, the number of PCR cycles could in future be reduced so the reaction is terminated during its linear phase. This would improve the quantification process and provide clearer information on the relative abundance of secondary alleles in the mixture. However, as the primary aim of this study was to analyse data pertaining to the major genotype, sensitivity was important and this optimisation was not performed. However, a large number of duplicate PCR and sequencer assays confirmed that the most abundant allele was consistently identified. This is supported by the observation that the secondary allele generally corresponded to only about three quarters of the peak area of the primary allele (Figure 3.15.), and therefore the primary and secondary alleles could, in general, be easily distinguished.

The markers developed in this study were used to address basic questions about the population genetics of *T. annulata*. Future applications of this set of markers could include experimental situations such as genetic crosses. For example, an animal could be co-infected with distinct clonal genotypes of *T. annulata* and once infection reaches the piroplasm stage, tick nymphs could be infected by feeding on the animal. Following meiosis in the tick gut (Schein *et al.* 1975) and clonal expansion in the tick salivary glands, sporozoite stabilates may be created and used to establish *in vitro* cell lines from which parasite clones may be derived. Individual clones may then be analysed to determine whether they represent parental or recombinant genotypes and this data may be used to construct a genetic map. Such an approach may be used to empirically demonstrate genetic exchange and also to investigate the occurrence of self-fertilisation in *T. annulata*. The markers may also prove particularly useful in transmission studies and in vaccine trials

for genotyping of challenge and breakthrough strains of *T. annulata*. For example, vaccine trials involve immunising a group of cattle with a particular parasite genotype, whether an attenuated cell line or a subunit vaccine. To test the efficacy of the vaccine, the cattle are then challenged with homologous and heterologous strains of the parasite. The micro- and mini-satellite markers may be used to differentiate between homologous and various heterologous strains that may breakthrough and this may reveal whether any conferred immunity was strain-specific. Additionally, development of an extensive panel of markers to provide complete and even coverage of the genome would also be of interest and may allow the identification of regions of the genome under selective pressure. For example in *Plasmodium*, it has been shown that alleles associated with drug resistance may be identified through a micro-satellite based mapping approach (Anderson 2004). Neutral variations at micro-satellite loci, which are linked to the beneficial mutation at the locus under selection are also affected by a selective sweep, by way of a phenomenon termed 'hitchhiking'. To explain, in a neutrally evolving population, allelic variation may be observed over the entire genome (Figure 6.1.(i)). However, natural populations commonly experience demographic alterations such as population expansions or contractions. A temporary reduction in effective population size is termed a 'bottleneck' and results in reduced variation over the entire genome (Figure 6.1.(ii)). When a selective pressure, such as drug usage, is applied to a population then beneficial mutations, which confer drug resistance, may increase in frequency until they become fixed in the population. Neutral loci that are genetically linked (i.e. physically close) are also affected and they also show a reduction in polymorphism (Figure 6.1.(iii)). Thus, the 'hitchhiking' effect is predicted to occur in regions flanking loci that are under selection and manifests itself as an area of localised reduction in genetic diversity and elevated linkage disequilibrium. This has been demonstrated around the *dhfr* gene in pyrimethamine-resistant isolates of *P. falciparum* collected in East Africa (Pearce *et al.* 2005). With the *T. annulata* genome sequence now fully assembled, a comprehensive screen may now be conducted to identify a large panel of markers that conforms to a step-wise model of mutation and provides extensive coverage of all four chromosomes.

6.3. Genetic exchange and population structure

The major findings discussed above indicate a large amount of diversity within *T. annulata* populations not only within a locality but also within a single host. As every isolate analysed in the field study comprises multiple genotypes, ticks feeding on these cattle are highly likely to be co-infected. Although a principal genotype can be identified in the

Figure 6.1. Partitioning of chromosomal variability at three regions

Three different chromosomal regions are illustrated and in each region, micro-satellite loci are indicated by circles, squares and triangles with allelic states distinguished by filled and empty symbols.

(i) Neutral scenario

Allelic diversity is present in all three regions.

(ii) Population bottleneck

A population bottleneck results in a genome wide reduction in variability.

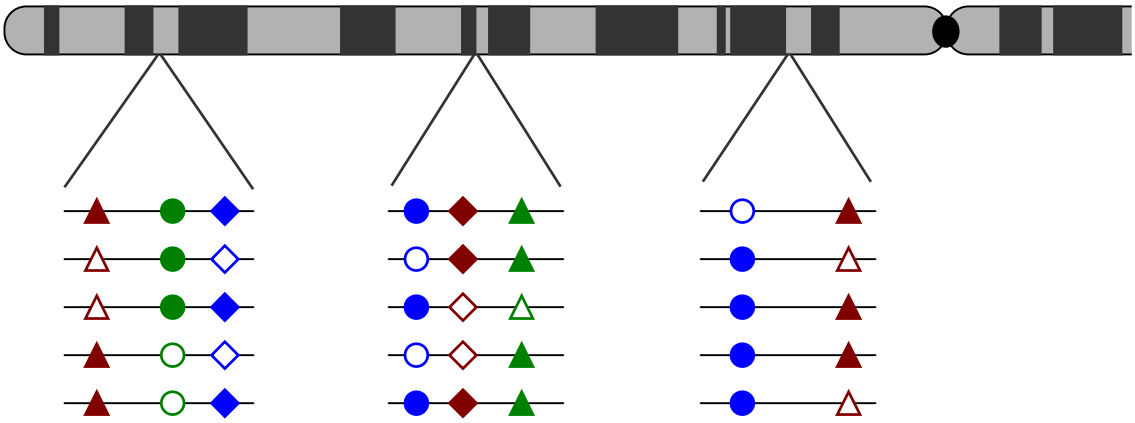
(iii) Recent selective sweep – the 'hitchhiking' effect

The central region was subjected to a recent selective sweep, resulting in reduced variability and linkage disequilibrium only at this locus.

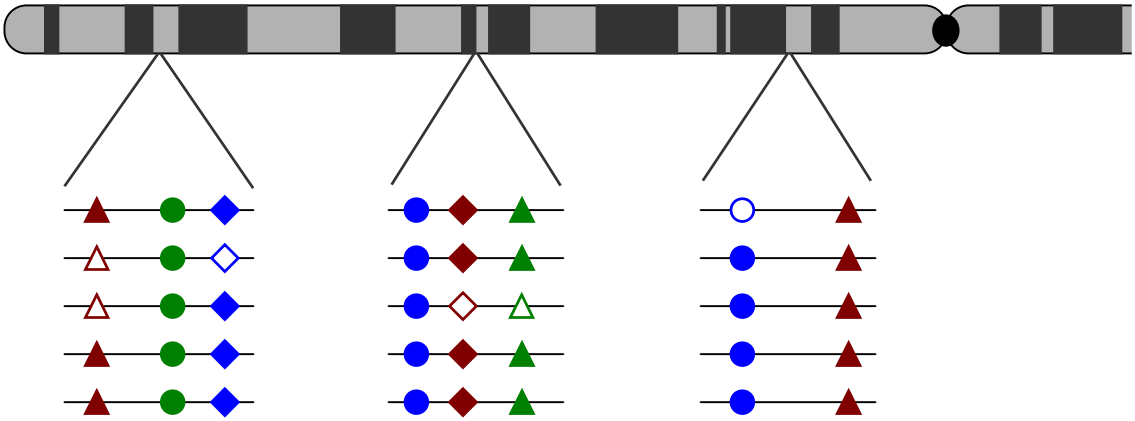
(Adapted from Schlötterer 2003)

Figure 6.1. Partitioning of chromosomal variability at three regions

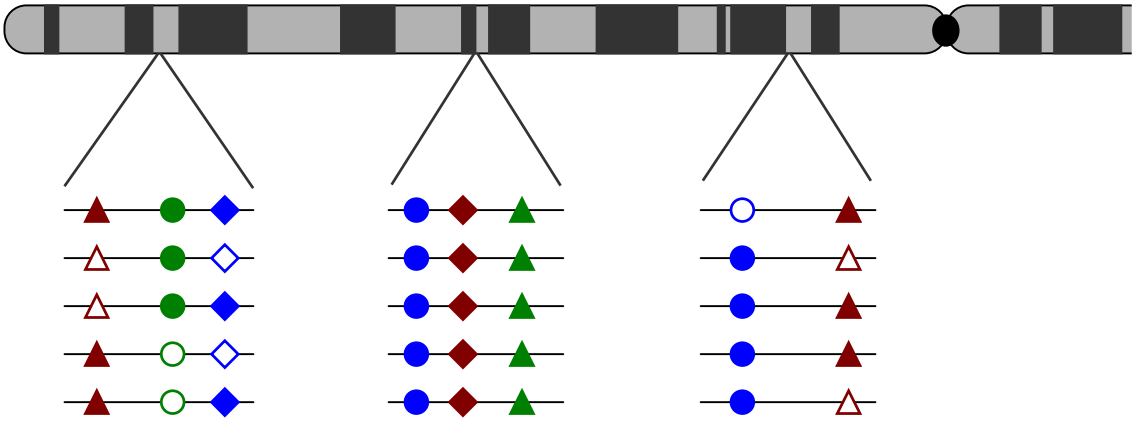
(i) Neutral scenario



(ii) Population bottleneck



(iii) Recent selective sweep – 'hitchhiking' effect



majority of field isolates, the abundance of the second allele can be approximated to 75 % of the primary allele and in many cases, the cumulative abundance of all the minor alleles may be greater than that of the primary allele. Therefore, underlying preconditions promoting genetic exchange are present, i.e. the co-infection of ticks with multiple genotypes at the same blood meal where a single genotype does not completely dominate.

The fact that out of 305 field samples analysed in Tunisia and Turkey, only one pair of isolates from the Aydın district were of identical genotype immediately implies that the parasite does not exhibit a simple clonal population structure. The relatively weak linkage disequilibrium identified in some regional populations and indicated by low values of the index of association were much lower than that observed in fully clonal bacterial populations (Smith *et al.* 1993; Haubold *et al.* 1998). In the preliminary analysis, the Tunisian population exhibited linkage equilibrium (LE), while in the second study it showed a slight departure over the country as a whole. This was attributed to one of the three farms in the village of El Hessiène being in slight linkage disequilibrium (LD), while all the other sampling sites exhibited LE. In Turkey, the sample size in the preliminary study was too small to draw any conclusions. However, in the second study, LD was detected over the whole sampling region, which broke down into LD in particular districts and villages and LE in others. With both LE and LD being detected within each country, the value of the I^S_A for each country as a whole (being the sum of the different regions) was lower than that detected in individual sampling areas. The limited LD detected in the study suggested a degree of relatedness between isolates from the same locality, indicating a slight departure from the population structure of panmixia. The various reasons for LD are fully discussed in Section 3.4.1., which include the presence of an epidemic population structure. How then should the population structure of *T. annulata* be interpreted? An epidemic population structure may be regarded as an intermediate between the extremes of panmixia and clonality, whereby a background level of sexual recombination is present that is masked by the vegetative expansion of a limited number of genotypes. Such a population structure was shown to occur in particular sub-types of the zoonotic pathogen *C. parvum* (Mallon *et al.* 2003b). A range of population structures have been shown to occur in *C. parvum* (*sensu lato*), ranging from clonality in human infective Type 1 infection (now termed *C. hominis*) to panmixia in the major sub-types of Type 2 (*C. parvum* (*sensu stricto*)) (Mallon *et al.* 2003a). Geographical and temporal sub-structuring were excluded from causing LD in a large collection of Type 2 isolates from different areas in Scotland (Mallon *et al.* 2003b), however in one particular area, while bovine isolates appeared panmictic, an epidemic population structure was provisionally

identified in a small group of human isolates. This was demonstrated by removing identical MLGs from the dataset and re-calculating the I^S_A, which dropped from 0.215 (indicating LD) to 0.001 (consistent with linkage equilibrium). It was concluded that the rapid expansion of a limited number of clonal genotypes in this population obscured the underlying high rate of genetic exchange. Similarly, an epidemic structure has also been demonstrated in populations of *T. parva* in Uganda, as when identical multilocus genotypes were treated as a single isolate, linkage disequilibrium was removed and panmixia identified (Oura *et al.* 2005). Interestingly in the *T. parva* study, a sub-group of isolates were removed before LD diminished in the population from Mbarara. Examination of the dendrogram representing the MLGs suggested that the Mbarara sub-group consisted of highly related parasite genotypes that shared a large proportion of alleles, but were substantially distinct from the rest of the population. Although not identical, this group of isolates led to the departure from LE. As in that analysis, if such a population is characterised as being epidemic, the implication is that the traditional method of removing absolutely identical isolates is too stringent when attempting to identify an epidemic population structure. This consideration stimulated a stratified, subtractive analysis to be performed on *T. annulata* populations in both Tunisia and Turkey (Table 3.8.). It was concluded that a small number of highly related genotypes were not responsible for causing LD in each country. In that analysis, it was necessary to remove almost half the isolates from each population before the remainder of isolates reverted to LE. This clearly indicated that *T. annulata* does not possess a generalised epidemic population structure. Although, it was suggested that the parasite may demonstrate such a structure at a micro-geographical scale (Section 3.4.2.), it may be concluded that the parasite population is essentially panmictic. Furthermore, when the Tunisian population is analysed on a site-by-site basis, LE (i.e. panmixia) is observed in three of the four sampling sites. Moreover, the Tunisian isolates show a low level of pair-wise similarity and the MLGs sharing the highest proportion of alleles do not necessarily originate from the same sampling site (Table 3.6.). This is highly consistent with the observation that LE is detected in the majority of sampling sites in Tunisia. In contrast, it is clear from the results presented in Table 3.7. that in Turkey there is a degree of relatedness between *T. annulata* isolates from the same district of origin and this may be taken as evidence for genetic sub-structuring of populations. As discussed in Section 3.4.1., sub-structuring between districts may explain LD over the country as a whole. Populations are partially sub-structured, which means that there is movement of parasites between populations so that when the populations are combined there is evidence for departure from panmixia. The flux of parasite genetic material between each district is not at a sufficient rate to

completely randomise allelic combinations in populations in separate districts in each generation. When one considers the ecology of the disease, this is intuitively correct. *T. annulata* is maintained in a purely cattle / tick transmission cycle with no paratenic host or *ex vivo* component to the life-cycle. Therefore gene flow between areas must be the result of movement of either the host or the vector. Compared to an airborne vector such as the anophelene mosquito, which may fly or be carried by the wind for considerable distances (Service 1997), the tick is likely to migrate over a very restricted range. Logically, cattle would be predicted to be the major vehicle for transporting *T. annulata* genotypes over moderate and large distances. Panmixia is an 'ideal' population structure and is defined as one where all individuals are potential mating partners. This requires that there are no restrictions, either genetic or spatial upon mating and as such all individuals in a population can recombine with each other. It is the ability of individuals to move within their natural range and thus mate with other members of the population that provides the necessary preconditions for random mating. Consequently the free movement of *T. annulata* would effectively be dependent on the free movement of cattle. Free movement of cattle over large distances, i.e. ranching, is not a feature of animal husbandry in either Tunisia or Turkey, and animals are kept exclusively on farms. Thus, panmixia of *T. annulata* could only really exist within a very limited area, perhaps within a single farm or village. This raises an apparent paradox - why do isolates from single farms or villages show a relatively high level of identity between isolates when panmixia may be operating at this scale? In a particular locality where the population exhibits a panmictic structure, alleles should be randomly reshuffled. However, notwithstanding novel mutation, alleles must pre-exist in the population before they can be reshuffled in a new generation. This raises a related question, in areas where linkage disequilibrium is (relatively) high, such as the district of Nazilli, is there actually a limited number of alleles in circulation? This can be determined, to an extent, by the results in Table 3.5., where the mean number of alleles present in each sample is presented (i.e. overall / country / district / village). The standard deviation for these values in all cases is large, due to the high level of variance exhibited across all ten loci, which in part may be attributed to varying levels of polymorphism between each marker. For instance, an average of 32.60 alleles at each locus is observed across the entire Turkish sample, whereas in the village of Nazilli the figure drops to 13.00, with a similar reduction in other districts, thus suggesting that across the markers, not all alleles are observed within an individual district. However, only the most abundant MLG / allele has been analysed in this study and, to an extent, the number of alleles identified positively correlates with the sample size. It is possible that alleles apparently absent in a particular district are present as minor components within each isolate. Although the

number of secondary 'peaks' in each isolate was recorded, these were not actually scored for genetic analysis. Scoring of these secondary alleles was performed in the preliminary study (Chapter Two), which comprised a smaller number of samples with a lower multiplicity of infection in comparison to the extensive field study (Chapter Three). Indeed, this full allelic profile was required for cluster analysis to differentiate between Tunisian and Turkish isolates. Scoring of secondary alleles in the extensive field study was not performed because it would have generated an extensive, unwieldy dataset that could only be interpreted by making a number of unacceptable assumptions. For example, it would have been difficult to calculate allele frequencies using this information, and the identification of secondary MLGs would have made too many assumptions. The only meaningful inference would have been identification of the total allelic range present in a given sample of isolates.

Linkage disequilibrium is detected within districts and villages (Table 3.5.), therefore, regardless of whether a large or small number of different alleles are present within the sample, alleles at pairs of loci are not distributed randomly and associations exist. Consequently, the findings of this study are inconsistent with panmixia and may be analogous with the situation described in *P. falciparum*. Similar to *T. annulata*, there is an obligate sexual cycle, however there is strong evidence of inbreeding and self-fertilisation in natural populations. The population genetic structure of *P. falciparum* has been shown to be predominantly clonal in regions of low transmission and this has been associated with inbreeding (Anderson *et al.* 2000a). In contrast, areas with high transmission intensity are believed to exhibit a panmictic population structure (Tibayrenc *et al.* 1990; Anderson *et al.* 2000a; Urdaneta *et al.* 2001). However several studies have indicated that inbreeding may also be a feature of areas of high transmission intensity (Paul *et al.* 1995; Razakandrainibe *et al.* 2005) and linkage disequilibrium has been identified in two populations from the Republic of the Congo, where transmission is high (Durand *et al.* 2003). In a recent micro-satellite study based in Kenya, inbreeding in areas of high infectivity was investigated in detail by examining the products of meiosis, i.e. individual oocysts, in the mid-gut of the mosquito vector (Razakandrainibe *et al.* 2005). A large discrepancy was identified between observed and expected heterozygosity of *P. falciparum* genotypes, indicating a considerable amount of inbreeding. Additionally, allele association was observed with pair-wise combinations of seven loci, which were distributed across five chromosomes. A proportion of inbreeding was directly attributed to selfing while a proportion was attributed to non-random distribution of genotypes among mosquito guts. That is to say, inbreeding was detectable within each mosquito that harboured two or more genotypes. Inbreeding,

including self-fertilisation, could certainly explain the linkage disequilibrium identified in *T. annulata* populations from localities analysed in this study. With one exception, parasite isolates possessing identical genotypes were not encountered and therefore there is no direct evidence of clonal replication and self-fertilisation. However, as discussed in Section 2.4.5., parasite DNA preparations which are known to contain a proportion of common genotypes may only share a proportion of predominant alleles. This was demonstrated when MLGs for homologous piroplasm and cell line preparation were compared demonstrating that between five and seven predominant alleles were shared in the most heterogeneous samples (Figure 2.6.). In addition, although the immediate products of self-fertilisation were not evident (i.e. multiple identical clonal genotypes), the isolates genotyped in this study may represent descendents from the offspring generated by earlier rounds of this form of mating.

Zygote formation in *T. annulata* takes place in the tick gut following which meiosis is presumed to occur (Schein and Friedhoff 1978). It must be borne in mind that the tick itself may represents an obligate 'bottleneck' in the *T. annulata* population. Following co-infection during feeding, it is unknown (a) how many genotypes an individual tick will ingest, (b) if there is selection in the tick (perhaps by bovine antibody), (c) how many zygotes may be formed, (d) how many kinetes migrate to the salivary glands, (e) how many kinetes establish in the salivary glands to permit parasite amplification. It would be of great interest to experimentally investigate the population dynamics in the vector in the first instance, and then proceed to determine the genotypes infecting individual ticks in an endemic area. Such a study would present considerable technical challenges. In order to measure observed heterozygosity, it would be necessary to genotype from the diploid stage of the organism. PCR amplification from a single zygote or kinete may be necessary since these are the only diploid stages in the parasite life-cycle. Although, presumably a single acinus represents the clonal expansion of a single kinete, therefore this stage of the parasite would also reflect the diploid product of meiosis. Additionally, sampling from a single acinus may provide a better representation of the genotypes present in the field than that obtained by analysing cattle blood samples. In this study, it was demonstrated that on average the secondary allele represented 75 % of the predominant allele at a model locus (TS5, Figure 3.18.). Therefore, it could be argued that the sampling method used in this study was problematic and that a 'true' estimate of the genotypes present was not being obtained.

It is clear from the findings of this study that the genetic structure of *T. annulata* in field populations is complex and lies close to panmixia, with a varying but limited level of linkage disequilibrium observed in different sampling areas. To definitively answer many of the questions raised here a structured epidemiological analysis must be undertaken. Basic aspects of the biology of the parasite and its interaction with both the vector and host need to be studied before definitive population structures can be defined. However, it is clear that sexual recombination occurs to an extent and plays a significant role in generating diversity in natural populations of *T. annulata*. In addition to the effects observed at a population genetic level, sexual recombination is also likely to promote polymorphism at the neutral marker loci used in this study. Polymorphism at non-neutral loci, such as antigen genes, is considered to be under the influence of selection. Investigating diversifying selection as the result of selective processes was the other major theme of this study.

6.4. *In silico* identification of antigens in the genome of *T. annulata*

The publication of the annotated genome sequence of *T. annulata* presented a novel opportunity to identify candidates for inclusion in a subunit vaccine. Starting from a genome of almost 4,000 CDS, a filtration process was employed which identified a subset of five putative antigen genes. Following an extensive allelic diversity study of two of these genes, TA13810 (*mero1*) was identified as a merozoite vaccine candidate antigen.

The first step in this process was the identification of a subset of predicted CDS that encoded *T. annulata* surface proteins, which may be exposed to the host immune system. While within the bovine host, the vast majority of the parasite's life is spent in an intra-cellular location and hence shielded from the immune response. Only transiently can the parasite be found in the host blood stream - during initial infection (the sporozoite) and during migration from the leucocyte to the erythrocyte (the merozoite). The merozoite stage was the focus of this study, primarily because EST expression data was available, allowing 20 % of predicted CDS to be the focus of analysis, although the sporozoite may prove to be an equally fruitful target for future investigations. In future, it would be interesting to determine whether high interspecies $d_N d_S$ is also associated with sporozoite surface proteins, in order to support or refute the hypothesis that transient extra-cellular exposure is associated with positive immune selection. A potential strategy for identifying sporozoite targets would be to identify proteins with a signal sequence and GPI anchor, but

without expression data. Northern Blotting or reverse transcription-polymerase chain reaction (RT-PCR) could be undertaken to screen a panel of these genes to detect if mRNA is present in sporozoite preparations. However, it may be prudent to wait for sporozoite EST data to be generated or for the results of future micro-array analyses. Although the macroschizont-infected leucocyte is the principal target for natural, protective immunity (Ahmed and Mehlhorn 1999), to date, no parasite-encoded proteins have been identified on the surface of the *T. annulata* macroschizont-infected cell. Recent work in *T. parva* has identified antigens that are presented on the infected cell surface as peptides by MHC Class I molecules (Graham *et al.* 2006). However, it would be impossible to identify a meaningful set of antigen genes encoding such peptides in a genome-wide screen using the bioinformatic software that is currently available. Although proprietary applications can predict Class I epitopes, the vast majority of algorithms are based on Human Leucocyte Antigen (HLA) binding predictions (Zhang *et al.* 2005; Doytchinova *et al.* 2006) and the value of such software in identifying bovine epitopes is unclear. Although limited software is available for predicting binding of the bovine MHC I allele - A20, http://bimas.dcrt.nih.gov/molbio/hla_bind/, (Parker *et al.* 1994), it is not designed for genome-wide screening applications. Hence, with present bioinformatic resources, the identification of genes encoding peptides presented during the macroschizont stage of the life-cycle was not feasible. However, it was possible to predict genes that are secreted by the macroschizont and are targeted for degradation in the host cell cytoplasm, which is a necessary pre-condition for presentation on MHC molecules. These represented genes with macroschizont expression data, a signal peptide and single or multiple PEST sequences and corresponded to the TashAT and SVSP gene families as discussed in the previous section. Similar to predicted merozoite antigens, these genes exhibited relatively high inter-species $d_N d_S$, indicating positive selection.

The ever-increasing number of complete genome-sequencing projects has facilitated several genome-wide scans for positive selection in a variety of species. Many of these studies have been conducted on the human genome, seeking to discover functionally important loci (Biswas and Akey 2006). Comparative $d_N d_S$ studies have been carried out between chimpanzees and man identifying genes involved in the immune response, sensory perception and gametogenesis (Nielsen *et al.* 2005). To date, in *T. annulata* the molecules involved in mechanisms such as gametogenesis have not been the focus of research, and many are likely to be annotated as hypothetical proteins. It is possible that these and other as yet unclassified genes are under the influence of positive selective pressures, unrelated to the bovine immune response. For example, one may predict that a

large number of genes are stage-specifically expressed as the parasite develops within the tick. A proportion of these genes may be genus or species-specific and thus share little similarity with annotated genes in current sequence databases. Future micro-array studies have the potential to identify these genes and it may be possible to determine whether tick-stage genes show evidence of positive selection.

The inter-species comparative analysis presented in this thesis was used to identify classes of gene under the influence of positive selection in *T. annulata*. It was principally used as 'proof of concept', demonstrating that putative merozoite surface molecules, identified by bioinformatic motif prediction were driven to diversify. While similar bioinformatic motif prediction approaches had been used in other protozoa (Bhatia *et al.* 2004), the availability of the genome of a closely related species allowed preliminary analysis of diversity among various classes of gene. A clear correlation was established between the presence of a signal peptide and an increasing level of $d_N d_S$ - around 30 % of genes with a $d_N d_S$ greater than 0.30 possessed a signal peptide in contrast to around 6 % of genes with $d_N d_S$ below 0.05 (Figure 4.2.). Most importantly, putative merozoite surface genes, possessing both a signal sequence and a GPI anchor were shown to have an elevated $d_N d_S$ in comparison to putative internal merozoite genes, which lacked either or both features (Figure 4.5.). This was taken as preliminary evidence of positive selective pressure acting on merozoite surface molecules in the two *Theileria* species.

The genome was bioinformatically screened to identify those particular genes, which encode a signal peptide, GPI anchor and have merozoite expression data in order to identify a panel of putative merozoite surface antigens. The analysis identified eight genes, which were ranked in order of descending $d_N d_S$ (Table 4.1.). The eight genes fell neatly into two classes, the first of which comprised three conserved proteins. These genes were generally large (up to 4.7 kb), possessed introns and exhibited low $d_N d_S$. They had previously been characterised in non-*Theileria* species as enzymes and a putative ion channel and were therefore considered unlikely to be antigenic. In contrast, the second class represented five promising antigen candidates. These were, in general, smaller in length, encoded by a single exon and critically possessed high $d_N d_S$ values. Encouragingly, one of these was the major *T. annulata* merozoite antigen, TaMS1, while three others had orthologues previously identified as antigens in other *Theileria* species. Sequence diversity of *TaMS1* had already been studied in detail (Katzer *et al.* 1998; Gubbels *et al.* 2000b) and this molecule has already undergone trials to assess its potential as a component of a sub-unit vaccine (d'Oliveira *et al.* 1997; Boulter *et al.* 1998). The two

highest rankings $d_N d_S$ genes were orthologues of *T. parva* sporozoite antigens, representing components of secretory organelles that are understood to be deployed following leucocyte invasion (Shaw 1997). The first is an orthologue of the *T. parva* microsphere antigen, p150 (Skilton *et al.* 1998) while the second is orthologous to the microneme-rhoptry antigen, p104 (Iams *et al.* 1990a; Iams *et al.* 1990b). Rhoptries and microspheres discharge their contents after entry into the host cell (Shaw 1997) and as such the two proteins are not found on the surface of the sporozoite. Related proteins of other apicomplexan parasites have, nevertheless, been shown to stimulate the immune response. For example, in *Toxoplasma gondii* where secretory organelles called dense granules discharge immediately following host cell invasion (Carruthers and Sibley 1997), B and T cell epitopes have been identified in GRA proteins and they have induced protection against experimental infection (Cesbron-Delauw 1994).

In *T. annulata* the p150 orthologue has EST data corresponding only to the merozoite stage, however in *T. parva*, RT-PCR and Northern blotting suggested expression is principally in the sporozoite and macroschizont stages (Skilton *et al.* 1998). No evidence was found for expression of p150 in the *T. parva* merozoite, consistent with the fact that microspheres are not present in this stage (Shaw 1997); consequently in *T. annulata*, the p150 orthologue should not be considered a microsphere protein. Again, this does not necessarily negate its function as a potential antigen. For example, merozoite organelle proteins are major components of a protein fraction of *Babesia bovis* which has been shown to induce protective immunity in cattle (Goodger *et al.* 1992), although re-invasion of erythrocytes by merozoites is a feature of this parasite species (Levine 1988). The high inter-species $d_N d_S$ of p150 also suggests that the molecule has undergone considerable divergence from *T. parva*, particularly at the amino acid level. Although a GPI anchor was indicated in the *T. annulata* orthologue, neither a transmembrane domain nor a GPI anchor was identified in the *T. parva* p150. Taken together, the data suggest that this molecule has adapted both in its stage of expression and its sub-cellular location. This suggests that although clearly closely related, the genes in *T. annulata* and *T. parva* may not perform the same function in each species and therefore should perhaps not be described as orthologues. Polymorphism had been observed among stocks of *T. parva* particularly at the C-terminus of p150 and it would have been very interesting to analyse diversity in isolates of *T. annulata*. p104 has been characterised to a lesser extent than p150, although it has been demonstrated to associate with the microneme/rhoptry complexes of the sporozoite using immuno-electron microscopy (Iams *et al.* 1990b). Hydrophobic stretches of amino acids located in the first and last 19 residues of the open reading frame, suggest

that the gene encodes a signal peptide at the N-terminus and a GPI-anchor at the C-terminus, similar to the *T. annulata* orthologue. EST data indicates that the orthologue of p104 is expressed in the macroschizont and merozoite of *T. annulata*, suggesting expression may be maintained from the sporozoite stage onwards. Bovine anti-sporozoite anti-serum generated using the infection and treatment protocol and sera from cattle subject to natural challenge were both shown to react with p104 (Iams *et al.* 1990b). This indicates that p104 encodes epitopes, which are exposed to and targeted by the bovine immune system. Due to the large size of both genes (2.9 kb and 2.7 kb), it was decided that time and resources were best directed towards analysing the two smaller novel antigens. However, the *T. annulata* 'orthologues' of p104 and p150 are worthy of further research in order to assess their antigenicity, the significance of their relatively high $d_N d_S$ and their stage-specific expression pattern.

It is unlikely that more information could have been easily gleaned from the inter-species comparison. In general, only limited regions of antigens are exposed to the bovine immune system and the effects of positive selection are focussed on small areas of the encoding gene. Analysing mean $d_N d_S$ across entire genes may therefore be considered a relatively insensitive technique in that a background of purifying selection may mask the presence of small tracts of positively selected nucleotides. Consequently, use of $d_N d_S$ was restricted to group analysis in order to identify trends between classes of gene. Fully utilising the genomic sequence diversity between the two species would have required additional bioinformatic techniques. For example, a sliding windows approach may be undertaken over an entire chromosome to identify genes possessing foci of positive selection. In future, techniques such as this may be employed to reveal a wealth of data from this information-rich resource. In this study, the power of the filtration technique was related to the process of identifying genes, which possessed relatively rare signature motifs. With only five genes identified from the entire genome, some *bona fide* merozoite antigens may have been discarded. However, for the purpose of this analysis, the degree of filtration was adequate because only a limited number of genes could be investigated using the allelic-sequencing methodology. Genuine surface antigens may have been overlooked through (1) lack of recognition of the signal motif, (2) inability to recognise proteins secreted by the non-classical pathway, (3) variation in the GPI motif of *T. annulata* and the model used by the prediction software and (4) insufficient EST data. The impact of the first three criteria is difficult to assess. Most motif prediction programs are based on models constructed using higher eukaryotes and it is possible subtle differences may exist within protozoan sequences. Although the sensitivity of the bioinformatic filtration

protocol is difficult to quantify, the results demonstrate that it is highly specific. Of the five putative antigen genes identified in the study (Table 4.1.), four had previously been identified as encoding antigens in *Theileria* species, although only one of these, *TaMS1* had been characterised in *T. annulata*. The putative antigen genes TA13810 (*mero1*) and TA20615 (*mero2*) were fed forward into an allelic diversity study to determine whether they showed evidence of positive selection within the *T. annulata* population.

6.5. Characterising putative antigen genes

The sequencing component of the study investigated whether diversity among alleles of *mero1* and *mero2* could be attributed to bovine immune selection. This involved determining allelic sequences for both genes across a panel of isolates from both Tunisia and Turkey, which showed *mero1* as a polymorphic antigen with potential as a vaccine candidate. In contrast, *mero2* was shown to be virtually invariant at the amino acid level both within and between populations, although polymorphism was evident at the nucleotide level. This indicated that mero2 is strongly under the influence of purifying selection and it does not encode a polymorphic antigen. Although slightly disappointing, this finding underlines the usefulness of the sequencing study, while casting an element of doubt on the genomic filtration methodology. From the genomic analysis alone, *mero2* in some ways appears to be the better candidate of the two. For example, it shows more interspecies diversity than *mero1* both at the DNA and protein level, it has a higher rate of $d_N d_S$ and it is stage-specifically expressed solely in the merozoite (Table 5.1.). The underlying reasons for this gene's monomorphism, despite its bioinformatic signature, are discussed in Section 5.4.5. Although, the study demonstrates that *mero2* is not diverse within *T. annulata* populations, it does not definitively show that it is not antigenic. Were *mero2* capable of inducing a protective immune response, it would be a very promising vaccine candidate indeed as the evidence of purifying selection suggests it has an essential function. However, this is inconsistent with the premise that antigen proteins are by their nature polymorphic and therefore antigenicity of the protein encoded by *mero2* is considered to be unlikely.

It is worthwhile considering how the allele sequencing protocol may be improved and what adaptations should be incorporated in future studies. As a form of experimental control, parasite-encoded host nuclear genes were included in the analysis. With field populations of *T. annulata* being extremely heterogeneous, it may be argued that extensive polymorphism would inevitably be encountered at any locus under investigation. To an extent, inclusion of the host nuclear genes was intended to disprove the notion that all

genes with high $d_N d_S$ are necessarily polymorphic within *T. annulata*. Though limited polymorphism was documented in the host nuclear genes, it is acknowledged they were not an ideal choice of control. For example, utilising putative internal merozoite genes encoding hypothetical proteins would have been more stringent. This would have permitted a direct comparison between internal (shielded) and external (exposed) proteins with identical stage-specific expression. Such genes were not included in the study for several reasons – (1) *mero2* functioned as an internal control, demonstrating that purifying selection could be detected across heterogeneous parasite genotypes, (2) studies in other organisms such as *Plasmodium* have supported the validity of multiple allelic sequencing, and therefore the technique is already considered valid and credible, without the need for a control and (3) a disproportionate amount of time and resources would have been spent generating a large amount of relatively uninformative data.

$d_N d_S$ analysis and the tests of neutrality used the allelic sequences of *T. annulata* alone, while the McDonald-Kreitman analysis required the use of an out-species. *T. parva* was selected since it is closely related to *T. annulata* and orthologous sequences for all six of the genes under study were available (i.e. merozoite, host nuclear and SVSP genes). In addition, several *T. buffeli / orientalis* sequences were used when investigating diversity of *mero1* and significantly lower neutrality indices were obtained using these genes than when using the *T. parva* sequence. The results presented in Table 5.10. demonstrate that the *T. parva* orthologue is more closely related to the *T. annulata* gene than are the other three orthologues. It possesses half as many non-synonymous and around quarter as many synonymous sites compared to the *T. buffeli / orientalis* sequences. In other words, positive selection is only demonstrated in comparison with the most closely related species. This raises the question – is *T. parva* the most useful comparator for this analysis? It is possible that the close relative of *T. annulata*, *T. lestoquardi* would be a better choice. However, it is difficult to predict how informative a very similar parasite would be. If divergence is limited and the immune response is very similar to that elicited by *T. annulata*, then sampling from such a species would be effectively similar to sampling an additional *T. annulata* allele. However, micro- and mini-satellite genotyping suggests an appreciable level of differentiation between the species (Section 2.4.3.). Furthermore, *T. lestoquardi* has adapted to the sheep and consequently the immune response to which it is exposed may not be identical and therefore selective pressures may be different. It would be possible to evaluate the efficacy of a single genotype of *T. lestoquardi* for the McDonald-Kreitman test by using the six genes analysed in this study. Whether

informative or not, a fully sequenced genome of *T. lestoquardi* would undoubtedly be a useful resource for future inter-genomic studies.

Further analyses could have been performed in this study had the allele sequencing protocol been modified slightly. In *Plasmodium*, low F_{ST} values have been associated with genes which are under balancing selection (Conway and Polley 2002). That is to say, in separate populations a limited number of identical alleles are maintained and this has been attributed to frequency dependent selection. F_{ST} measures the reduction in heterozygosity when comparing the entire population to defined sub-populations. Hence, this measurement can only be made if heterozygosity can be accurately estimated in each sub-population. By sequencing individual clones from the Tunisian population, it would have been possible to estimate heterozygosity at each locus. The Turkish sequences were derived from four highly heterogeneous parasite mixtures and although multiple sequences were generated from each sample, some were unique and some were identical. It was therefore difficult to assess how allelic diversity measured across the unique alleles in the Turkish population related to actual heterozygosity. Consequently, the use of F_{ST} analysis was precluded since heterozygosity could not be accurately estimated. This is unfortunate, since the F_{ST} values for the micro- and mini-satellite markers determined in Chapters Two and Three may have been ideal to serve as neutral controls. Nevertheless, the sequencing methodology and the statistical tests used in this study were capable of identifying the novel merozoite antigen gene, *mero1*.

6.6. A novel merozoite vaccine candidate, *mero1*

During the course of this study *mero1* was suggested to be under the influence of several different forms of selection. Although above average, the inter-species $d_N d_S$ value was the lowest of the six genes under study and nucleotide and amino acid identities between *T. annulata* and *T. parva* were the highest (Table 5.1.). This implied the gene was comparatively well conserved between the species and that the general structure of the gene was likely to be constrained. This was echoed at the allelic level within *T. annulata*, when only two size variants differing by 3 bp were observed among almost 60 alleles. In marked contrast, host nuclear and SVSP genes showed considerable length polymorphism with several sites of insertion and deletion. Signal peptide and GPI motifs were also well conserved within *mero1*, demonstrating that a proportion of the gene is under the influence of purifying selection. However, it was the overall DNA and protein diversity data that provided the most interesting results with this gene. In both Tunisia and Turkey, *mero1* displayed the highest level of nucleotide diversity, in sharp contrast to limited diversity

between the species. When the nature of this diversity was further investigated, *mero1* was shown to be under the influence of both positive and balancing selection. Positive selection was demonstrated by an excess of non-synonymous substitutions found within the *mero1* population of *T. annulata* when compared against the orthologous gene in several different *Theileria* species (Table 5.10.). $d_N d_S$ analysis demonstrated that these sites were not uniformly distributed across the gene, but were found in clusters, the most striking of which centred around codon 150 (Figure 5.9.(i)). Evidence of balancing selection was provided by the Tajima's *D* and Fu & Li's *D* and *F* tests (Figure 5.10. and Table 5.12.) whereby multiple alleles at this locus are maintained in a population. Taken together, this suggests a 'tug-of-war' is taking place in *mero1* between various selective pressures. That is to say, there is a general background of purifying selection, which acts to conserve the general structure of the encoded protein. Concurrently, the gene is driven to diversify probably by the host immune response; however this diversity is constrained through frequency-dependent (balancing) selection, which acts to maintain alleles in the population. Therefore, although a large number of different alleles have been identified in the population, the total number of allelic 'types' may be effectively limited. This complex pattern of behaviour suggests the gene may (a) encode an antigenic protein and (b) be a good vaccine candidate. Independent corroborative evidence for *mero1* encoding an antigen is provided by the orthologous gene in *T. sergenti*, which encodes a polymorphic 23 kDa piroplasm antigen (Zhuang *et al.* 1995; Sako *et al.* 1999). Whether the gene is a good vaccine candidate is more open to debate. The principal arguments in support of its suitability are outlined in Section 5.4.5. However, two important factors should be considered –

1. Extensive amino acid polymorphism may preclude efficacy of mero1. It is often stated that highly polymorphic proteins are unsuitable as vaccine targets because immunising with a single or limited number of allelic forms may fail to protect against those genotypes possessing variant alleles. While a valid consideration, it must be stressed that *mero1* was not identified on the basis of 'excessive' polymorphism. Both $d_N d_S$ analysis and the McDonald-Kreitman test are based on the rate of amino acid altering nucleotide substitutions to silent ones. For example the *Plasmodium* liver stage antigen gene, *lsa-1* (Zhu and Hollingdale 1991) shows elevated $d_N d_S$; minimal diversity is encountered at the nucleotide level, however this diversity is biased towards encoding variant amino acids. Put another way, high $d_N d_S$ does not necessarily equate with extensive polymorphism. Moreover, balancing selection, which constrains allelic diversity within the populations has been shown to influence diversity of *mero1*. It should also be

considered that while populations of *T. annulata* are highly heterogeneous, with multiple genotypes harboured in every infected individual (Chapter Three), protective immunity against field populations can be elicited following cell line immunisation with a limited number of genotypes. This suggests that the enormous polymorphism exhibited between different *T. annulata* genotypes may not lead to evasion of a broadly cross-protective immune response. It may be argued that natural and induced immunity are non-specific or that the protective mechanisms involved are independent of the merozoite stage of infection. However, this leads to the question of - what pressure drives the selection of amino acid substitution in this molecule within *T. annulata*? It may be hypothesised that merozoite antigen diversity is not a result of selection by the bovine host, but rather it may reflect ancestral polymorphism present before *T. annulata* established in cattle. In buffalo, the natural host of this genus of parasites, *T. parva* has been shown to be maintained in a carrier state for many years (Schreuder *et al.* 1977). Merozoite antigen diversity may reflect evasion or attempted evasion of the buffalo immune response and may therefore reflect polymorphism of an ancestral *Theileria* parasite in a different host species. Moreover, if (hypothetically) erythrocyte infection was perpetuated by constant re-invasion by merozoites (a feature of other *Theileria* species (Kawamoto *et al.* 1990)) several generations of the merozoite would have been exposed to positive selective pressure within each life-cycle of the parasite. This is analogous to the theory that ancestral polymorphism explains the abundance of synonymous substitutions in *mero2*, as discussed in Section 5.4.6. Although it is possible that the immune response of the buffalo is in part directed against the merozoite, it is likely that, similar to cattle, the main target is the macroschizont-infected leucocyte.

2. The merozoite is not the principal target of natural immunity. In *T. annulata* the macroschizont is considered to be the life-cycle stage which is most likely to be targeted by the natural immune response, with $CD8^+$ cytotoxic T-cells presumed to recognise MHC class I associated antigen on infected leucocytes (Ahmed *et al.* 1989). Until a decade ago, there was no evidence that anti-merozoite antibodies were involved in immunity generated from either cell line vaccination or natural infection (Irvin 1985; Hall 1988). However, evidence now exists that merozoite surface proteins may offer a degree of protection. Immune calf sera was shown to recognise recombinant TaMS1 protein by Western Blotting (d'Oliveira *et al.* 1996) and immunisation with such proteins elicited a protective immune response (d'Oliveira *et al.* 1997). However, these two finding do not necessarily directly correlate – i.e. antibody may be present, but it may not necessarily confer protection. Therefore there is still no clear evidence in *T. annulata* that antibody

can protect against the merozoite stage *in vivo* or *in vitro*. However, in *T. sergenti*, *in vivo* merozoite invasion was inhibited following transfusion with a monoclonal antibody raised to the orthologue of TaMS1 (Tanaka *et al.* 1990). These studies clearly indicate the potential for merozoite surface proteins to elicit an immune response, which even if quite different from that conferred by natural infection or cell line vaccination may be of value in disease control. With infected leucocytes being responsible for the major pathology of the disease, immunity conferred by *mero1* may simply ameliorate the condition of an infected animal rather than prevent onset of clinical signs. Additionally, it may afford the infected animal a better chance to successfully combat the disease in concert with the natural immune response, although productivity losses may still occur. If immune selection is indeed taking place in the bovine, then reduced merozoite invasion would limit the number of piroplasm-infected erythrocytes in the circulation and this may in turn have a transmission blocking (or reducing) effect. However, it should be considered that anti-merozoite antibodies may not function to prevent merozoite invasion in cattle and that bovine immune selection may actually be occurring in the tick on the ingested piroplasm stage. Nevertheless, a transmission blocking effect may still be predicted to take place, whereby the parasite fails to establish in the tick.

In conclusion, *mero1* should be the subject of further study to assess its suitability for inclusion in a sub-unit vaccine. Initially, this may take the form of further bioinformatic studies. For example, proprietary software may be used to predict the presence and location of bovine B and T cell epitopes on the gene. If such epitopes were demonstrated to co-locate with clusters of high $d_N d_S$, this would provide further evidence that the gene is diversifying due to bovine immune selection. Inevitably, the point will come when *in silico* prediction must give way to *in vitro* experiments. The first step in such an event may be the generation of recombinant mero1 protein. Similar to the situation in TaMS1, this could be screened against sera of immune animals to determine whether it is recognised by the bovine immune response. Ultimately, *in vivo* studies may be indicated, whereby DNA, peptide or protein immunogens based on *mero1* are tested for their ability to protect against homologous and heterologous challenge in the field.

Overall, this study represents a natural progression in the direction of research on *T. annulata*, whereby the published genome is used as a resource for generating hypotheses, which are then tested at the bench. This post-genomic era should continue to inform on fundamental aspects of the biology of the parasite and stimulate new avenues of research, both at the theoretical and applied level. Continued studies on the generation of

genomic diversity will provide information on how the *Theileria* have adapted to different host-species environments and evolved to evade their immune responses. Moreover, genes and biochemical pathways, which have co-evolved to exploit different biological niches may be investigated in order to identify both novel therapeutic drug targets and the genes involved in emerging drug resistance. In particular, the need to identify those antigens, which can elicit broad protective immunity against *T. annulata* has been highlighted in this thesis. Hopefully, in the not too distant future, bioinformatic techniques will have evolved sufficiently to achieve this goal.

References

Abdel-Rahman, M.S., Fahmy, M.M., & Aggour, M.G. (1998). Trials for control of ixodid ticks using pheromone acaricide tick decoys, *J. Egypt. Soc Parasitol.*, **28**, 551-557.

Adamson, R. & Hall, R. (1997). A role for matrix metalloproteinases in the pathology and attenuation of *Theileria annulata* infections, *Parasitol. Today*, **13**, 390-393.

Adamson, R., Logan, M., Kinnaird, J., Langsley, G., & 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.

Adamson, R.E. & Hall, F.R. (1996). Matrix metalloproteinases mediate the metastatic phenotype of *Theileria annulata*-transformed cells, *Parasitology*, **113 (Pt 5)**, 449-455.

Ahmed, J.S., Diesing, L., Oechtering, H., Ouhelli, H., & Schein, E. (1988). The role of antibodies in immunity against *Theileria annulata* infection in cattle, *Zentralbl. Bakteriol. Mikrobiol. Hyg.*, **267**, 425-431.

Ahmed, J.S. & Mehlhorn, H. (1999). Review: the cellular basis of the immunity to and immunopathogenesis of tropical theileriosis, *Parasitol. Res.*, **85**, 539-549.

Ahmed, J.S., Rothert, M., Steuber, S., & Schein, E. (1989). *In vitro* proliferative and cytotoxic responses of PBL from *Theileria annulata*-immune cattle, *Zentralbl. Veterinarmed.*, **36**, 584-592.

Aktas, M., Dumanli, N., & Angin, M. (2004). Cattle infestation by *Hyalomma* ticks and prevalence of *Theileria* in *Hyalomma* species in the east of Turkey, *Vet. Parasitol.*, **119**, 1-8.

Allan, E. & Wren, B.W. (2003). Genes to genetic immunization: identification of bacterial vaccine candidates, *Methods*, **31**, 193-198.

Anderson, T.J. (2004). Mapping drug resistance genes in *Plasmodium falciparum* by genome-wide association, *Curr. Drug Targets. Infect. Disord.*, **4**, 65-78.

Anderson, T.J. & Day, K.P. (2000). Geographical structure and sequence evolution as inferred from the *Plasmodium falciparum* S-antigen locus, *Mol. Biochem. Parasitol.*, **106**, 321-326.

Anderson, T.J., Haubold, B., Williams, J.T., Estrada-Franco, J.G., Richardson, L., Mollinedo, R., Bockarie, M., Mokili, J., Mharakurwa, S., French, N., Whitworth, J., Velez, I.D., Brockman, A.H., Nosten, F., Ferreira, M.U., & Day, K.P. (2000a). Microsatellite markers reveal a spectrum of population structures in the malaria parasite *Plasmodium falciparum*, *Mol. Biol. Evol.*, **17**, 1467-1482.

Anderson, T.J., Su, X.Z., Roddam, A., & Day, K.P. (2000b). Complex mutations in a high proportion of microsatellite loci from the protozoan parasite *Plasmodium falciparum*, *Mol. Ecol.*, **9**, 1599-1608.

Andersson, S.G. & Kurland, C.G. (1990). Codon preferences in free-living microorganisms, *Microbiol. Rev.*, **54**, 198-210.

Bakheit, M.A. & 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.

Bakheit, M.A., Schnittger, L., Salih, D.A., Boguslawski, K., Beyer, D., Fadl, M., & Ahmed, J.S. (2004). Application of the recombinant *Theileria annulata* surface protein in an indirect ELISA for the diagnosis of tropical theileriosis, *Parasitol. Res.*, **92**, 299-302.

Barnes, W.M. (1994). PCR amplification of up to 35-kb DNA with high fidelity and high yield from lambda bacteriophage templates, *Proc. Natl. Acad. Sci. USA*, **91**, 2216-2220.

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

Barry, J.D. & McCulloch, R. (2001). Antigenic variation in trypanosomes: enhanced phenotypic variation in a eukaryotic parasite, *Adv. Parasitol.*, **49**, 1-70.

Baton, L.A. & Ranford-Cartwright, L.C. (2005). Spreading the seeds of million-murdering death: metamorphoses of malaria in the mosquito, *Trends Parasitol.*, **21**, 573-580.

Baylis, H.A., Megson, A., Brown, C.G., Wilkie, G.F., & Hall, R. (1992). *Theileria annulata*-infected cells produce abundant proteases whose activity is reduced by long-term cell culture, *Parasitology*, **105 (Pt 3)**, 417-423.

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

Ben Miled, L. (1993). *Population Diversity in Theileria annulata in Tunisia,* PhD thesis, University of Edinburgh.

Ben Miled, L., Dellagi, K., Bernardi, G., Melrose, T.R., Darghouth, M., Bouattour, A., Kinnaird, J., Shiels, B., Tait, A., & Brown, C.G. (1994). Genomic and phenotypic diversity of Tunisian *Theileria annulata* isolates, *Parasitology*, **108 (Pt 1)**, 51-60.

Bendtsen, J.D., Nielsen, H., von Heijne, G., & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0, *J. Mol. Biol.*, **340**, 783-795.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences, *Nucleic Acids Res.*, **27**, 573-580.

Bettencourt, A., Franca, C., & Borges, J. (1907). Un cas de piroplasmose bacilliforme chez le daim, *Institut Royal de Bacteriologie*, **1**, 341-363.

Bhatia, V., Sinha, M., Luxon, B., & Garg, N. (2004). Utility of the *Trypanosoma cruzi* sequence database for identification of potential vaccine candidates by *in silico* and *in vitro* screening, *Infect. Immun.*, **72**, 6245-6254.

Bishop, R., Geysen, D., Skilton, R., Odongo, D., Nene, V., Allsopp, B., Mbogo, S., Spooner, P., & Morzaria, S. (2002). Genomic polymorphism, sexual recombination and molecular epidemiology of *Theileria parva*, in *Theileria*, Kluwer Academic, Dordrecht, 23-40.

Bishop, R., Geysen, D., Spooner, P., Skilton, R., Nene, V., Dolan, T., & Morzaria, S. (2001). Molecular and immunological characterisation of *Theileria parva* stocks which are components of the 'Muguga cocktail' used for vaccination against East Coast fever in cattle, *Vet. Parasitol.*, **94**, 227-237.

Bishop, R., Morzaria, S., & Gobright, E. (1998). Linkage of two distinct AT-rich minisatellites at multiple loci in the genome of *Theileria parva*, *Gene*, **216**, 245-254.

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

Bishop, R., Nene, V., Staeyert, J., Rowlands, J., Nyanjui, J., Osaso, J., Morzaria, S., & Musoke, A. (2003). Immunity to East Coast fever in cattle induced by a polypeptide fragment of the major surface coat protein of *Theileria parva* sporozoites, *Vaccine*, **21**, 1205-1212.

Bishop, R.P., Sohanpal, B.K., & Morzaria, S.P. (1994a). Cloning and characterisation of a repetitive DNA sequence from *Theileria mutans*: application as a species-specific probe, *Parasitol. Res.*, **80**, 33-41.

Bishop, R.P., Spooner, P.R., Kanhai, G.K., Kiarie, J., Latif, A.A., Hove, T., Masaka, S., & Dolan, T.T. (1994b). Molecular characterization of *Theileria* parasites: application to the epidemiology of theileriosis in Zimbabwe, *Parasitology*, **109 (Pt 5)**, 573-581.

Biswas, S. & Akey, J.M. (2006). Genomic insights into positive selection, *Trends Genet.*, **22**, 437-446.

Blood, C.H., Sasse, J., Brodt, P., & Zetter, B.R. (1988). Identification of a tumor cell receptor for VGVAPG, an elastin-derived chemotactic peptide, *J. Cell Biol.*, **107**, 1987-1993.

Blouin, M.S., Parsons, M., Lacaille, V., & Lotz, S. (1996). Use of microsatellite loci to classify individuals by relatedness, *Mol. Ecol.*, **5**, 393-401.

Blythe, J.E., Surentheran, T., & Preiser, P.R. (2004). STEVOR--a multifunctional protein?, *Mol. Biochem. Parasitol.*, **134**, 11-15.

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

Boulter, N., Brown, D., Wilkie, G., Williamson, S., Kirvar, E., Knight, P., Glass, E., Campbell, J., Morzaria, S., Nene, V., Musoke, A., d'Oliveira, C., Gubbels, M.J., Jongejan, F., & Hall, R. (1999). Evaluation of recombinant sporozoite antigen SPAG-1 as a vaccine candidate against *Theileria annulata* by the use of different delivery systems, *Trop. Med. Int. Health*, **4**, A71-A77.

Boulter, N., Knight, P.A., Hunt, P.D., Hennessey, E.S., Katzer, F., Tait, A., Williamson, S., Brown, D., Baylis, H.A., & Hall, R. (1994). *Theileria annulata* sporozoite surface antigen (SPAG-1) contains neutralizing determinants in the C terminus, *Parasite Immunol.*, **16**, 97-104.

Boulter, N.R., Brown, C.G., Kirvar, E., Glass, E., Campbell, J., Morzaria, S., Nene, V., Musoke, A., d'Oliveira, C., Gubbels, M.J., Jongejan, F., & Hall, F.R. (1998). Different vaccine strategies used to protect against *Theileria annulata*, *Ann. N. Y. Acad. Sci.*, **849**, 234-246.

Boulter, N.R., Glass, E.J., Knight, P.A., Bell-Sakyi, L., Brown, C.G., & Hall, R. (1995). *Theileria annulata* sporozoite antigen fused to hepatitis B core antigen used in a vaccination trial, *Vaccine*, **13**, 1152-1160.

Bowcock, A.M., Ruiz-Linares, A., Tomfohrde, J., Minch, E., Kidd, J.R., & Cavalli-Sforza, L.L. (1994). High resolution of human evolutionary trees with polymorphic microsatellites, *Nature*, **368**, 455-457.

Boyle, J.P., Rajasekar, B., Saeij, J.P., Ajioka, J.W., Berriman, M., Paulsen, I., Roos, D.S., Sibley, L.D., White, M.W., & Boothroyd, J.C. (2006). Just one cross appears capable of dramatically altering the population biology of a eukaryotic pathogen like *Toxoplasma gondii*, *Proc. Natl. Acad. Sci. USA*, **103**, 10514-10519.

Britten, R.J. (1998). Precise sequence complementarity between yeast chromosome ends and two classes of just-subtelomeric sequences, *Proc. Natl. Acad. Sci. USA*, **95**, 5906-5912.

Brown, C.G. (1981). Application of *in vitro* techniques to vaccination against theileriosis, in *Advances in the control of theileriosis*, A.D. Irvin, M.P. Cunningham, & A.S. Young, Martinus Nijhoff, The Hague, 104-119.

Brown, D.J., Campbell, J.D., Russell, G.C., Hopkins, J., & Glass, E.J. (1995). T cell activation by *Theileria annulata*-infected macrophages correlates with cytokine production, *Clin. Exp. Immunol.*, **102**, 507-514.

Campbell, J.D., Brown, D.J., Nichani, A.K., Howie, S.E., Spooner, R.L., & Glass, E.J. (1997a). A non-protective T helper 1 response against the intra-macrophage protozoan *Theileria annulata*, *Clin. Exp. Immunol.*, **108**, 463-470.

Campbell, J.D., Nichani, A.K., Brown, D.J., Howie, S.E., Spooner, R.L., & Glass, E.J. (1997b). Parasite-mediated steps in immune response failure during primary *Theileria annulata* infection, *Trop. Anim. Health Prod.*, **29**, 133S-135S.

Carruthers, V.B. & Sibley, L.D. (1997). Sequential protein secretion from three distinct organelles of *Toxoplasma gondii* accompanies invasion of human fibroblasts, *Eur. J. Cell Biol.*, **73**, 114-123.

Cesbron-Delauw, M.F. (1994). Dense-granule organelles of *Toxoplasma gondii*: their role in the host-parasite relationship, *Parasitol. Today*, **10**, 293-296.

Chanda, I., Pan, A., & Dutta, C. (2005). Proteome composition in *Plasmodium falciparum*: higher usage of GC-rich nonsynonymous codons in highly expressed genes, *J. Mol. Evol.*, **61**, 513-523.

Chansiri, K., Kawazu, S., Kamio, T., Terada, Y., Fujisaki, K., Philippe, H., & Sarataphan, N. (1999). Molecular phylogenetic studies on *Theileria* parasites based on small subunit ribosomal RNA gene sequences, *Vet. Parasitol.*, **83**, 99-105.

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

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry, C.E., III, Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J.E., Taylor, K., Whitehead, S., & Barrell, B.G. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence, *Nature*, **393**, 537-544.

Conrad, P.A., Iams, K., Brown, W.C., Sohanpal, B., & ole-MoiYoi, O.K. (1987). DNA probes detect genomic diversity in *Theileria parva* stocks, *Mol. Biochem. Parasitol.*, **25**, 213-226.

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

Conway, D.J., Machado, R.L., Singh, B., Dessert, P., Mikes, Z.S., Povoa, M.M., Oduola, A.M., & Roper, C. (2001). Extreme geographical fixation of variation in the *Plasmodium falciparum* gamete surface protein gene *Pfs48/45* compared with microsatellite loci, *Mol. Biochem. Parasitol.*, **115**, 145-156.

Conway, D.J. & Polley, S.D. (2002). Measuring immune selection, *Parasitology*, **125**, S3-16.

Cowman, A.F. & Crabb, B.S. (2005). Revealing the molecular determinants of gender in malaria parasites, *Cell*, **121**, 659-660.

d'Oliveira, C., Feenstra, A., Vos, H., Osterhaus, A.D., Shiels, B.R., 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**, 1796-1804.

d'Oliveira, C., Tijhaar, E.J., Shiels, B.R., van der, W.M., & Jongejan, F. (1996). Expression of genes encoding two major *Theileria annulata* merozoite surface antigens in *Escherichia coli* and a *Salmonella typhimurium* aroA vaccine strain, *Gene*, **172**, 33-39.

d'Oliveira, C., van der, W.M., Habela, M.A., Jacquiet, P., & Jongejan, F. (1995). Detection of *Theileria annulata* in blood samples of carrier cattle by PCR, *J. Clin. Microbiol.*, **33**, 2665-2669.

Darghouth, M.A., Ben Miled, L., Bouattour, A., Melrose, T.R., Brown, C.G., & Kilani, M. (1996a). A preliminary study on the attenuation of Tunisian schizont-infected cell lines of *Theileria annulata*, *Parasitol. Res.*, **82**, 647-655.

Darghouth, M.A., Bouattour, A., & Kilan, M. (1999). Tropical theileriosis in Tunisia: epidemiology and control, *Parassitologia*, **41 (S1)**, 33-36.

Darghouth, M.E., Bouattour, A., Ben Miled, L., Kilani, M., & Brown, C.G. (1996b). Epidemiology of tropical theileriosis (*Theileria annulata* infection of cattle) in an endemic region of Tunisia: characterisation of endemicity states, *Vet. Parasitol.*, **65**, 199-211.

Darghouth, M.E., Bouattour, A., Ben Miled, L., & Sassi, L. (1996c). Diagnosis of *Theileria annulata* infection of cattle in Tunisia: comparison of serology and blood smears, *Vet. Res.*, **27**, 613-621.

De Groot, A.S. & Rappuoli, R. (2004). Genome-derived vaccines, *Expert. Rev. Vaccines.*, **3**, 59-76.

de Kok, J.B., d'Oliveira, C., & Jongejan, F. (1993). Detection of the protozoan parasite *Theileria annulata* in *Hyalomma* ticks by the polymerase chain reaction, *Exp. Appl. Acarol.*, **17**, 839-846.

de Vos, A.J., Bessenger, R., & Banting, L.F. (1981). *Theileria? taurotragi*: a probable agent of bovine cerebral theileriosis, *Onderstepoort J. Vet. Res.*, **48**, 177-178.

Debrauwere, H., Gendrel, C.G., Lechat, S., & Dutreix, M. (1997). Differences and similarities between various tandem repeat sequences: minisatellites and microsatellites, *Biochimie*, **79**, 577-586.

Dickson, J. & Shiels, B.R. (1993). Antigenic diversity of a major merozoite surface molecule in *Theileria annulata*, *Mol. Biochem. Parasitol.*, **57**, 55-64.

Dolan, T.T., Teale, A.J., Stagg, D.A., Kemp, S.J., Cowan, K.M., Young, A.S., Groocock, C.M., Leitch, B.L., Spooner, R.L., & Brown, C.G. (1984). A histocompatibility barrier to immunization against East Coast fever using *Theileria parva*-infected lymphoblastoid cell lines, *Parasite Immunol.*, **6**, 243-250.

Dong, H., Nilsson, L., & Kurland, C.G. (1996). Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates, *J. Mol. Biol.*, **260**, 649-663.

Doolan, D.L., Houghten, R.A., & Good, M.F. (1991). Location of human cytotoxic T cell epitopes within a polymorphic domain of the *Plasmodium falciparum* circumsporozoite protein, *Int. Immunol.*, **3**, 511-516.

Doytchinova, I.A., Guan, P., & Flower, D.R. (2006). EpiJen: a server for multistep T cell epitope prediction, *BMC. Bioinformatics.*, **7**, 131.

Dschunkowsky, E. & Luhs, J. (1904). Die piroplasmosen der rinder, *Centralblatt fur Bakteriologie, Parasitenkunde, Infektionskrankheit und Hygiene*, **1**, 486-492.

Dumanli, N., Aktas, M., Cetinkaya, B., Cakmak, A., Koroglu, E., Saki, C.E., Erdogmus, Z., Nalbantoglu, S., Ongor, H., Simsek, S., Karahan, M., & Altay, K. (2005). Prevalence and distribution of tropical theileriosis in eastern Turkey, *Vet. Parasitol.*, **127**, 9-15.

Durand, P., Michalakis, Y., Cestier, S., Oury, B., Leclerc, M.C., Tibayrenc, M., & Renaud, F. (2003). Significant linkage disequilibrium and high genetic diversity in a population of *Plasmodium falciparum* from an area (Republic of the Congo) highly endemic for malaria, *Am. J. Trop. Med. Hyg.*, **68**, 345-349.

Endo, T., Ikeo, K., & Gojobori, T. (1996). Large-scale search for genes on which positive selection may operate, *Mol. Biol. Evol.*, **13**, 685-690.

Enright, A.J., Van Dongen, S., & Ouzounis, C.A. (2002). An efficient algorithm for large-scale detection of protein families, *Nucleic Acids Res.*, **30**, 1575-1584.

Escalante, A.A., Lal, A.A., & Ayala, F.J. (1998). Genetic polymorphism and natural selection in the malaria parasite *Plasmodium falciparum*, *Genetics*, **149**, 189-202.

Estrada-Pena, A. (2003). Climate change decreases habitat suitability for some tick species (Acari: Ixodidae) in South Africa, *Onderstepoort J. Vet. Res.*, **70**, 79-93.

Fawcett, D.W., Doxsey, S., Stagg, D.A., & Young, A.S. (1982). The entry of sporozoites of *Theileria parva* into bovine lymphocytes *in vitro*. Electron microscopic observations, *Eur. J. Cell Biol.*, **27**, 10-21.

Fayer, R., Morgan, U., & Upton, S.J. (2000). Epidemiology of *Cryptosporidium*: transmission, detection and identification, *Int. J. Parasitol.*, **30**, 1305-1322.

Fields, S., Kohara, Y., & Lockhart, D.J. (1999). Functional genomics, *Proc. Natl. Acad. Sci. USA*, **96**, 8825-8826.

Fivaz, B.H., Norval, R.A., & Lawrence, J.A. (1989). Transmission of *Theileria parva bovis* (Boleni strain) to cattle resistant to the brown ear tick *Rhipicephalus appendiculatus* (Neumann), *Trop. Anim. Health Prod.*, **21**, 129-134.

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

Flach, E.J., Ouhelli, H., Waddington, D., Oudich, M., & Spooner, R.L. (1995). Factors influencing the transmission and incidence of tropical theileriosis (*Theileria annulata* infection of cattle) in Morocco, *Vet. Parasitol.*, **59**, 177-188.

Forbes, S.H., Hogg, J.T., Buchanan, F.C., Crawford, A.M., & Allendorf, F.W. (1995). Microsatellite evolution in congeneric mammals: domestic and bighorn sheep, *Mol. Biol. Evol.*, **12**, 1106-1113.

Forsyth, L.M., Minns, F.C., Kirvar, E., Adamson, R.E., Hall, F.R., McOrist, S., Brown, C.G., & Preston, P.M. (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.

Fu, Y.X. & Li, W.H. (1993). Statistical tests of neutrality of mutations, *Genetics*, **133**, 693-709.

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

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**, 498-511.

Geysen, D., Bazarusanga, T., Brandt, J., & Dolan, T.T. (2004). An unusual mosaic structure of the PIM gene of *Theileria parva* and its relationship to allelic diversity, *Mol. Biochem. Parasitol.*, **133**, 163-173.

Geysen, D., Bishop, R., Skilton, R., Dolan, T.T., & Morzaria, S. (1999). Molecular epidemiology of *Theileria parva* in the field, *Trop. Med. Int. Health*, **4**, A21-A27.

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

Gill, B.S., Bansal, G.C., Bhattacharyulu, Y., Kaur, D., & Singh, A. (1980). Immunological relationship between strains of *Theileria annulata* Dschunkowsky and Luhs 1904, *Res. Vet. Sci.*, **29**, 93-97.

Gill, B.S., Bhattacharyulu, Y., & Kaur, D. (1976). Immunisation against bovine tropical theileriasis (*Theileria annulata* infection), *Res. Vet. Sci.*, **21**, 146-149.

Gill, P., Jeffreys, A.J., & Werrett, D.J. (1985). Forensic application of DNA 'fingerprints', *Nature*, **318**, 577-579.

Glascodine, J., Tetley, L., Tait, A., Brown, D., & Shiels, B. (1990). Developmental expression of a *Theileria annulata* merozoite surface antigen, *Mol. Biochem. Parasitol.*, **40**, 105-112.

Glass, E.J., Innes, E.A., Spooner, R.L., & 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., & Brown, C.G. (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.

Good, M.F., Pombo, D., Quakyi, I.A., Riley, E.M., Houghten, R.A., Menon, A., Alling, D.W., Berzofsky, J.A., & Miller, L.H. (1988). Human T-cell recognition of the circumsporozoite protein of *Plasmodium falciparum*: immunodominant T-cell domains map to the polymorphic regions of the molecule, *Proc. Natl. Acad. Sci. USA*, **85**, 1199-1203.

Goodger, B.V., Waltisbuhl, D.J., Commins, M.A., & Wright, I.G. (1992). *Babesia bovis*: dextran sulphate as an adjuvant for and precipitant of protective immunogens, *Int. J. Parasitol.*, **22**, 465-469.

Graham, S.P., Pelle, R., Honda, Y., Mwangi, D.M., Tonukari, N.J., Yamage, M., Glew, E.J., de Villiers, E.P., Shah, T., Bishop, R., Abuya, E., Awino, E., Gachanja, J., Luyai, A.E., Mbwika, F., Muthiani, A.M., Ndegwa, D.M., Njahira, M., Nyanjui, J.K., Onono, F.O., Osaso, J., Saya, R.M., Wildmann, C., Fraser, C.M., Maudlin, I., Gardner, M.J., Morzaria, S.P., Loosmore, S., Gilbert, S.C., Audonnet, J.C., van der, B.P., Nene, V., & Taracha, E.L. (2006). *Theileria parva* candidate vaccine antigens recognized by immune bovine cytotoxic T lymphocytes, *Proc. Natl. Acad. Sci. USA*, **103**, 3286-3291.

Grassly, N.C. & Holmes, E.C. (1997). A likelihood method for the detection of selection and recombination using nucleotide sequences, *Mol. Biol. Evol.*, **14**, 239-247.

Gray, M.A. & Brown, C.G. (1981). *Advances in the control of theileriosis*.

Gray, M.A., Luckins, A.G., Rae, P.F., & Brown, C.G. (1980). Evaluation of an enzyme immunoassay for serodiagnosis of infections with *Theileria parva* and *T. annulata*, *Res. Vet. Sci*, **29**, 360-366.

Gubbels, M.J., d'Oliveira, C., & Jongejan, F. (2000a). 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., Katzer, F., Hide, G., Jongejan, F., & Shiels, B.R. (2000b). Generation of a mosaic pattern of diversity in the major merozoite-piroplasm surface antigen of *Theileria annulata*, *Mol. Biochem. Parasitol.*, **110**, 23-32.

Gubbels, M.J., Katzer, F., Shiels, B.R., & Jongejan, F. (2001). Study of *Theileria annulata* population structure during bovine infection and following transmission to ticks, *Parasitology*, **123**, 553-561.

Hall, F.R. (1988). Antigens and immunity in *Theileria annulata*, *Parasitol. Today*, **4**, 257-261.

Hall, R., Coggins, L., McKellar, S., Shiels, B., & Tait, A. (1990). Characterisation of an extrachromosomal DNA element from *Theileria annulata*, *Mol. Biochem. Parasitol.*, **38**, 253-260.

Hall, R., Hunt, P.D., Carrington, M., Simmons, D., Williamson, S., Mecham, R.P., & Tait, A. (1992). Mimicry of elastin repetitive motifs by *Theileria annulata* sporozoite surface antigen, *Mol. Biochem. Parasitol.*, **53**, 105-112.

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**, A78-A84.

Hammer, M.F., Blackmer, F., Garrigan, D., Nachman, M.W., & Wilder, J.A. (2003). Human population structure and its effects on sampling Y chromosome sequence variation, *Genetics*, **164**, 1495-1509.

Harr, B., Weiss, S., David, J.R., Brem, G., & Schlotterer, C. (1998). A microsatellite-based multilocus phylogeny of the *Drosophila melanogaster* species complex, *Curr. Biol.*, **8**, 1183-1186.

Hashemi-Fesharki, R. (1988). Control of *Theileria annulata* in Iran, *Parasitol. Today*, **4**, 36-40.

Hashemi-Fesharki, R. (1991). Prophylactic effect of schizont tissue culture vaccine against *Theileria annulata* infection in Iran (Conference Proceedings), National Dairy Development Board, Anand, India, 15-17.

Haubold, B. & Hudson, R.R. (2000). LIAN 3.0: detecting linkage disequilibrium in multilocus data. Linkage Analysis, *Bioinformatics.*, **16**, 847-848.

Haubold, B., Travisano, M., Rainey, P.B., & Hudson, R.R. (1998). Detecting linkage disequilibrium in bacterial populations, *Genetics*, **150**, 1341-1348.

Hedrick, P.W. (2005). Neutral Theory and Coalescence, in *Genetics of Populations*, 3rd edition, P.W. Hedrick, Jones and Bartlett, Sudbury, MA, 407-468.

Higgins, D.G. & Sharp, P.M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer, *Gene*, **73**, 237-244.

Hooshmand-Rad, P. (1985). The use of tissue culture attenuated live vaccine for *Theileria hirci*, *Dev. Biol. Stand.*, **62**, 119-127.

Hooshmand-Rad, P. & Hashemi-Fesharki, R. (1968). The effect of virulence on culture of *Theileria annulata* strains in lymphoid cells which have been cultured in suspension, *Arch. Inst. Razi*, **20**, 85-89.

Hudson, R.R., Kreitman, M., & Aguade, M. (1987). A test of neutral molecular evolution based on nucleotide data, *Genetics*, **116**, 153-159.

Hughes, A.L., Ota, T., & Nei, M. (1990). Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules, *Mol. Biol. Evol.*, **7**, 515-524.

Hughes, M.K. & Hughes, A.L. (1995). Natural selection on *Plasmodium* surface proteins, *Mol. Biochem. Parasitol.*, **71**, 99-113.

Iams, K.P., Hall, R., Webster, P., & Musoke, A.J. (1990a). Identification of lambda gt11 clones encoding the major antigenic determinants expressed by *Theileria parva* sporozoites, *Infect. Immun.*, **58**, 1828-1834.

Iams, K.P., Young, J.R., Nene, V., Desai, J., Webster, P., ole-MoiYoi, O.K., & Musoke, A.J. (1990b). Characterisation of the gene encoding a 104-kilodalton microneme-rhoptry protein of *Theileria parva*, *Mol. Biochem. Parasitol.*, **39**, 47-60.

Ilhan, T. (1999). *Diagnostic methods for epidemiological studies of tropical theileriosis (Theileria annulata infection of cattle),* PhD, University of Edinburgh.

Ilhan, T., Williamson, S., Kirvar, E., Shiels, B., & Brown, C.G. (1998). *Theileria annulata*: carrier state and immunity, *Ann. N. Y. Acad. Sci.*, **849**, 109-125.

Innes, E.A., Millar, P., Brown, C.G., & Spooner, R.L. (1989a). The development and specificity of cytotoxic cells in cattle immunized with autologous or allogeneic *Theileria annulata*-infected lymphoblastoid cell lines, *Parasite Immunol.*, **11**, 57-68.

Innes, E.A., Ouhelli, H., Oliver, R.A., Simpson, S.P., Brown, C.G., & Spooner, R.L. (1989b). The effect of MHC compatibility between parasite-infected cell line and recipient in immunization against tropical theileriosis, *Parasite Immunol.*, **11**, 47-56.

Irvin, A.D. (1985). Immunity in theileriosis, *Parasitol. Today*, **1**, 124-128.

Irvin, A.D. & Morrison, I.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.

Jaccard, P. (1908). *Bull. Soc. Vaudoise Sci. Nat.*, **44**, 223-270.

Jeffreys, A.J. (2005). Genetic fingerprinting, *Nat Med.*, **11**, 1035-1039.

Jeffreys, A.J., Royle, N.J., Wilson, V., & Wong, Z. (1988). Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA, *Nature*, **332**, 278-281.

Jeffreys, A.J., Wilson, V., & Thein, S.L. (1985a). Hypervariable 'minisatellite' regions in human DNA, *Nature*, **314**, 67-73.

Jeffreys, A.J., Wilson, V., & Thein, S.L. (1985b). Individual-specific 'fingerprints' of human DNA, *Nature*, **316**, 76-79.

Kanhai, G.K., Pegram, R.G., Hargreaves, S.K., Hove, T., & Dolan, T.T. (1997). Immunisation of cattle in Zimbabwe using *Theileria parva* (Boleni) without concurrent tetracycline therapy, *Trop. Anim. Health Prod.*, **29**, 92-98.

Katzer, F., Carrington, M., Knight, P., Williamson, S., Tait, A., Morrison, I.W., & Hall, R. (1994). Polymorphism of *SPAG-1*, a candidate antigen for inclusion in a sub-unit vaccine against *Theileria annulata*, *Mol. Biochem. Parasitol.*, **67**, 1-10.

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

Katzer, F., McKellar, S., Ferguson, M.A., d'Oliveira, C., & Shiels, B.R. (2002). A role for tertiary structure in the generation of antigenic diversity and molecular association of the *Tams1* polypeptide in *Theileria annulata*, *Mol. Biochem. Parasitol.*, **122**, 55-67.

Kaufmann, J. (1996). Parasites of Cattle, in *Parasitic Infections of Domestic Animals - A Diagnostic Manual*, Basel, 69-70.

Kaviratne, M., Khan, S.M., Jarra, W., & Preiser, P.R. (2002). Small variant STEVOR antigen is uniquely located within Maurer's clefts in *Plasmodium falciparum*-infected red blood cells, *Eukaryot. Cell*, **1**, 926-935.

Kawamoto, S., Takahashi, K., Kurosawa, T., Sonoda, M., & Onuma, M. (1990). Intraerythrocytic schizogony of *Theileria sergenti* in cattle, *Nippon Juigaku. Zasshi*, **52**, 1251-1259.

Kimura, M. (1969). The rate of molecular evolution considered from the standpoint of population genetics, *Proc. Natl. Acad. Sci. USA*, **63**, 1181-1188.

Kimura, M. (1979). The neutral theory of molecular evolution, *Sci. Am.*, **241**, 98-100, 102, 108.

Kirvar, E., Ilhan, T., Katzer, F., Wilkie, G., Hooshmand-Rad, P., & Brown, D. (1998). Detection of *Theileria lestoquardi* (*hirci*) in ticks, sheep, and goats using the polymerase chain reaction, *Ann. N. Y. Acad. Sci*, **849**, 52-62.

Knight, P., Musoke, A.J., Gachanja, J.N., Nene, V., Katzer, F., Boulter, N., Hall, R., Brown, C.G., Williamson, S., Kirvar, E., Bell-Sakyi, L., Hussain, K., & Tait, A. (1996). Conservation of neutralizing determinants between the sporozoite surface antigens of *Theileria annulata* and *Theileria parva*, *Exp. Parasitol.*, **82**, 229-241.

Kobayashi, N., Tamura, K., & Aotsuka, T. (1999). PCR error and molecular population genetics, *Biochem. Genet.*, **37**, 317-321.

Kocken, C.H., Jansen, J., Kaan, A.M., Beckers, P.J., Ponnudurai, T., Kaslow, D.C., Konings, R.N., & Schoenmakers, J.G. (1993). Cloning and expression of the gene coding for the transmission blocking target antigen *Pfs48/45* of *Plasmodium falciparum*, *Mol. Biochem. Parasitol.*, **61**, 59-68.

Kocken, C.H., Narum1 DL, Massougbodji, A., Ayivi, B., Dubbeld, M.A., van der, W.A., Conway, D.J., Sanni, A., & Thomas, A.W. (2000). Molecular characterisation of *Plasmodium reichenowi* apical membrane antigen-1 (*AMA-1*), comparison with *P. falciparum AMA-1*, and antibody-mediated inhibition of red cell invasion, *Mol. Biochem. Parasitol.*, **109**, 147-156.

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, *J. Mol. Biol.*, **305**, 567-580.

Kubota, S., Sugimoto, C., Kakuda, T., & Onuma, M. (1996). Analysis of immunodominant piroplasm surface antigen alleles in mixed populations of *Theileria sergenti* and *T. buffeli*, *Int. J. Parasitol.*, **26**, 741-747.

Kyes, S., Horrocks, P., & Newbold, C. (2001). Antigenic variation at the infected red cell surface in malaria, *Annu. Rev. Microbiol.*, **55**, 673-707.

Lawrence, J.A. (2006). Theileriosis of sheep and goats, in *Infectious Diseases of Livestock*, 2nd edition, vol. 1, J.A.W. Coetzer & R.C. Tustin, Oxford University Press, Oxford, 498-501.

Lawrence, J.A., Perry, B.D., & Williamson, S.M. (2006a). Corridor disease, in *Infectious Diseases of Livestock*, 2nd edition, vol. 1, J.A.W. Coetzer & R.C. Tustin, Oxford University Press, Oxford, 468-471.

Lawrence, J.A., Perry, B.D., & Williamson, S.M. (2006b). East Coast Fever, in *Infectious Diseases of Livestock*, 2nd edition, vol. 1, J.A.W. Coetzer & R.C. Tustin, Oxford University Press, Oxford, 448-467.

Lawrence, J.A. & Williamson, S.M. (2006a). *Theileria mutans* infection, in *Infectious Diseases of Livestock*, 2nd edition, vol. 1, J.A.W. Coetzer & R.C. Tustin, Oxford University Press, Oxford, 480-482.

Lawrence, J.A. & Williamson, S.M. (2006b). Turning sickness, in *Infectious Diseases of Livestock*, 2nd edition, vol. 1, J.A.W. Coetzer & R.C. Tustin, Oxford University Press, Oxford, 475-477.

Leemans, I., Brown, D., Hooshmand-Rad, P., Kirvar, E., & Uggla, A. (1999). Infectivity and cross-immunity studies of *Theileria lestoquardi* and *Theileria annulata* in sheep and cattle: I. *In vivo* responses, *Vet. Parasitol.*, **82**, 179-192.

Leemans, I., Hooshmand-Rad, P., & Uggla, A. (1997). The indirect fluorescent antibody test based on schizont antigen for study of the sheep parasite *Theileria lestoquardi*, *Vet. Parasitol.*, **69**, 9-18.

Levine, N.D. (1985). Apicomplexa: The Piroplasms, in *Veterinary Protozoology*, 291-328.

Levine, N.D. (1988). *The protozoan phylum apicomplexa,* CRC Press, Boca Raton, Florida.

Li, W.H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution, *J. Mol. Evol.*, **36**, 96-99.

Lobry, J.R. & Gautier, C. (1994). Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes, *Nucleic Acids Res.*, **22**, 3174-3180.

Louis, E.J. & Haber, J.E. (1990). Mitotic recombination among subtelomeric Y' repeats in *Saccharomyces cerevisiae*, *Genetics*, **124**, 547-559.

Lundberg, K.S., Shoemaker, D.D., Adams, M.W., Short, J.M., Sorge, J.A., & Mathur, E.J. (1991). High-fidelity amplification using a thermostable DNA polymerase isolated from *Pyrococcus furiosus*, *Gene*, **108**, 1-6.

Luqmani, Y.A., Mathew, M., Lobo, S., & Temmim, L. (1999). High-resolution electrophoretic DNA fragment analysis using Spreadex gels, *Anal. Biochem.*, **275**, 116-118.

MacLeod, A., Tweedie, A., Welburn, S.C., Maudlin, I., Turner, C.M., & Tait, A. (2000). Minisatellite marker analysis of *Trypanosoma brucei*: reconciliation of clonal, panmictic, and epidemic population genetic structures, *Proc. Natl. Acad. Sci. USA*, **97**, 13442-13447.

Maione, D., Margarit, I., Rinaudo, C.D., Masignani, V., Mora, M., Scarselli, M., Tettelin, H., Brettoni, C., Iacobini, E.T., Rosini, R., D'Agostino, N., Miorin, L., Buccato, S., Mariani, M., Galli, G., Nogarotto, R., Nardi, D., V, Vegni, F., Fraser, C., Mancuso, G., Teti, G., Madoff, L.C., Paoletti, L.C., Rappuoli, R., Kasper, D.L., Telford, J.L., & Grandi, G. (2005). Identification of a universal Group B streptococcus vaccine by multiple genome screen, *Science*, **309**, 148-150.

Mallon, M., MacLeod, A., Wastling, J., Smith, H., Reilly, B., & Tait, A. (2003a). Population Structures and the Role of Genetic Exchange in the Zoonotic Pathogen *Cryptosporidium parvum*, *J. Mol. Evol.*, **56**, 407-417.

Mallon, M.E., MacLeod, A., Wastling, J.M., Smith, H., & Tait, A. (2003b). Multilocus genotyping of *Cryptosporidium parvum* Type 2: population genetics and sub-structuring, *Infect. Genet. Evol.*, **3**, 207-218.

McDonald, J.H. & Kreitman, M. (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*, *Nature*, **351**, 652-654.

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

McKeever, D.J., Taracha, E.L., Innes, E.L., MacHugh, N.D., Awino, E., Goddeeris, B.M., & Morrison, W.I. (1994). Adoptive transfer of immunity to *Theileria parva* in the $CD8^+$ fraction of responding efferent lymph, *Proc. Natl. Acad. Sci U. S. A*, **91**, 1959-1963.

Mecham, R.P., Hinek, A., Entwistle, R., Wrenn, D.S., Griffin, G.L., & Senior, R.M. (1989). Elastin binds to a multifunctional 67-kilodalton peripheral membrane protein, *Biochemistry*, **28**, 3716-3722.

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

Miranda, J., Stumme, B., Beyer, D., Cruz, H., Oliva, A.G., Bakheit, M., Wicklein, D., Yin, H., Lou, J., Ahmed, J.S., & Seitzer, U. (2004). Identification of antigenic proteins of a *Theileria* species pathogenic for small ruminants in China recognized by antisera of infected animals, *Ann. N. Y. Acad. Sci.*, **1026**, 161-164.

Mishra, A.K., Sharma, N.N., & Viswanathan, C.B. (1993). Efficacy of Butalex in field cases of bovine theileriosis--short communication, *Acta Vet. Hung.*, **41**, 361-363.

Molano, A., Segura, C., Guzman, F., Lozada, D., & Patarroyo, M.E. (1992). In human malaria protective antibodies are directed mainly against the Lys-Glu ion pair within the Lys-Glu-Lys motif of the synthetic vaccine SPf 66, *Parasite Immunol.*, **14**, 111-124.

Moller, S., Croning, M.D., & Apweiler, R. (2001). Evaluation of methods for the prediction of membrane spanning regions, *Bioinformatics*, **17**, 646-653.

Montigiani, S., Falugi, F., Scarselli, M., Finco, O., Petracca, R., Galli, G., Mariani, M., Manetti, R., Agnusdei, M., Cevenini, R., Donati, M., Nogarotto, R., Norais, N., Garaguso, I., Nuti, S., Saletti, G., Rosa, D., Ratti, G., & Grandi, G. (2002). Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae*, *Infect. Immun.*, **70**, 368-379.

Morel, P.C. & Uilenberg, G. (1981). Sur la nomeclature de quelques Theileria (Sporozoa, Babesioidea) des ruminants domestiques, *Revue d'Elevage et de Medecine Veterinaire des Pays Tropiquaux*, **34**, 139-143.

Morrison, W.I. (1996). Influence of host and parasite genotypes on immunological control of Theileria parasites, *Parasitology*, **112**, S53-S66.

Morrison, W.I., Goddeeris, B.M., Teale, A.J., Groocock, C.M., Kemp, S.J., & Stagg, D.A. (1987). Cytotoxic T-cells elicited in cattle challenged with *Theileria parva* (Muguga): evidence for restriction by class I MHC determinants and parasite strain specificity, *Parasite Immunol.*, **9**, 563-578.

Morrison, W.I. & McKeever, D.J. (1998). Immunology of infections with *Theileria parva* in cattle, *Chem. Immunol.*, **70**, 163-185.

Morrison, W.I., Taracha, E.L., & McKeever, D.J. (1995). Theileriosis: progress towards vaccine development through understanding immune responses to the parasite, *Vet. Parasitol.*, **57**, 177-187.

Morzaria, S.P., Dolan, T.T., Norval, R.A., Bishop, R.P., & Spooner, P.R. (1995). Generation and characterization of cloned *Theileria parva* parasites, *Parasitology*, **111 (Pt 1)**, 39-49.

Moxon, E.R., Rainey, P.B., Nowak, M.A., & Lenski, R.E. (1994). Adaptive evolution of highly mutable loci in pathogenic bacteria, *Curr. Biol.*, **4**, 24-33.

Muhammed, S.I., Lauerman, L.H., Jr., & Johnson, L.W. (1975). Effect of humoral antibodies on the course of *Theileria parva* infection (East Coast fever) of cattle, *Am. J. Vet. Res.*, **36**, 399-402.

Musembi, S., Janoo, R., Sohanpal, B., Ochanda, H., ole-Moiyoi, O., Bishop, R., & Nene, V. (2000). Screening for *Theileria parva* secretory gene products by functional analysis in *Saccharomyces cerevisiae*, *Mol. Biochem. Parasitol.*, **109**, 81-87.

Musoke, A., Morzaria, S., Nkonge, C., Jones, E., & Nene, V. (1992). A recombinant sporozoite surface antigen of *Theileria parva* induces protection in cattle, *Proc. Natl. Acad. Sci. USA*, **89**, 514-518.

Musoke, A., Rowlands, J., Nene, V., Nyanjui, J., Katende, J., Spooner, P., Mwaura, S., Odongo, D., Nkonge, C., Mbogo, S., Bishop, R., & Morzaria, S. (2005). Subunit vaccine based on the *p67* major surface protein of *Theileria parva* sporozoites reduces severity of infection derived from field tick challenge, *Vaccine*, **23**, 3084-3095.

Musoke, A.J., Palmer, G.H., McElwain, T.F., Nene, V., & McKeever, D. (1996). Prospects for subunit vaccines against tick-borne diseases, *Br. Vet. J.*, **152**, 621-639.

Musto, H., Rodriguez-Maseda, H., & Bernardi, G. (1995). Compositional properties of nuclear genes from *Plasmodium falciparum*, *Gene*, **152**, 127-132.

Musto, H., Romero, H., Zavala, A., Jabbari, K., & Bernardi, G. (1999). Synonymous codon choices in the extremely GC-poor genome of *Plasmodium falciparum*: compositional constraints and translational selection, *J. Mol. Evol.*, **49**, 27-35.

Nacer, A., Berry, L., Slomianny, C., & Mattei, D. (2001). *Plasmodium falciparum* signal sequences: simply sequences or special signals?, *Int. J. Parasitol.*, **31**, 1371-1379.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals, *Genetics*, **89**, 583-590.

Nei, M. (1987). *Molecular Evolutionary Genetics,* Columbia University Press, New York.

Nei, M. & Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions, *Mol. Biol. Evol.*, **3**, 418-426.

Ngumi, P.N., Young, A.S., Lampard, D., Mining, S.K., Ndungu, S.G., Lesan, A.C., Williamson, S.M., Linyonyi, A., & Kariuki, D.P. (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.

Nielsen, H., Engelbrecht, J., Brunak, S., & von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites, *Protein Eng*, **10**, 1-6.

Nielsen, R., Bustamante, C., Clark, A.G., Glanowski, S., Sackton, T.B., Hubisz, M.J., Fledel-Alon, A., Tanenbaum, D.M., Civello, D., White, T.J., Sninsky, J., Adams, M.D., & Cargill, M. (2005). A scan for positively selected genes in the genomes of humans and chimpanzees, *PLoS. Biol.*, **3**, e170.

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

Nott, A., Meislin, S.H., & Moore, M.J. (2003). A quantitative analysis of intron effects on mammalian gene expression, *RNA.*, **9**, 607-617.

Odongo, D.O., Oura, C.A., Spooner, P.R., Kiara, H., Mburu, D., Hanotte, O.H., & Bishop, R.P. (2006). Linkage disequilibrium between alleles at highly polymorphic mini- and micro-satellite loci of *Theileria parva* isolated from cattle in three regions of Kenya, *Int. J. Parasitol.*

Omer, O.H., Haroun, E.M., Mahmoud, O.M., Abdel-Magied, E.M., El Malik, K.H., & Magzoub, M. (2003). Parasitological and clinico-pathological profiles in friesian cattle naturally infected with *Theileria annulata* in Saudi Arabia, *J. Vet. Med. B Infect. Dis. Vet. Public Health*, **50**, 200-203.

Ouhelli, H. (1985). *Theileriosis bovine a Theileria annulata (Dschunkowsky and Luhs, 1904). Recherche sur la biologie des vecteurs (Hyalomma spp.) et sur les interactions hôte-parasite,* Doctor of Science thesis, INP, Tououse.

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

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

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

Özkoc, U. & Pipano, E. (1981). Trials with cell culture vaccine against theileriosis in Turkey, in *Advances in the control of theileriosis*, A.D. Irvin, M.P. Cunningham, & A.S. Young, Martinus Nijhoff, The Hague, 256-258.

Page, R.D. (1996). TreeView: an application to display phylogenetic trees on personal computers, *Comput. Appl. Biosci.*, **12**, 357-358.

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

Palumbi, S.R. (1999). All males are not created equal: fertility differences depend on gamete recognition polymorphisms in sea urchins, *Proc. Natl. Acad. Sci. USA*, **96**, 12632-12637.

Park, S.D.E. (2001). *Trypanotolerance in West African Cattle and the Population Genetic Effects of Selection,* PhD thesis, University of Dublin.

Parker, K.C., Bednarek, M.A., & Coligan, J.E. (1994). Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains, *J. Immunol.*, **152**, 163-175.

Paul, R.E., Packer, M.J., Walmsley, M., Lagog, M., Ranford-Cartwright, L.C., Paru, R., & Day, K.P. (1995). Mating patterns in malaria parasite populations of Papua New Guinea, *Science*, **269**, 1709-1711.

Peakall, R. & Smouse, P.E. (2006). GENALEX6: genetic analysis in Excel. Population genetic software for teaching and research, *Molecular Ecology Notes*, **6**, 288-295.

Pearce, R., Malisa, A., Kachur, S.P., Barnes, K., Sharp, B., & Roper, C. (2005). Reduced variation around drug-resistant *dhfr* alleles in African *Plasmodium falciparum*, *Mol. Biol. Evol.*, **22**, 1834-1844.

Pegram, R.G., Wilson, D.D., & Hansen, J.W. (2000). Past and present national tick control programs. Why they succeed or fail, *Ann. N. Y. Acad. Sci.*, **916**, 546-554.

Pipano, E. (1974). Immunological aspects of *Theileria annulata* infection, *Bulletin de l'Office International des Epizooties*, **81**, 139-159.

Pipano, E. (1997). Vaccines against hemoparasitic diseases in Israel with special reference to quality assurance, *Trop. Anim. Health Prod.*, **29**, 86S-90S.

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

Pipano, E. & Shkap, V. (2006). *Theileria annulata* infection, in *Infectious Diseases of Livestock*, 2nd edition, vol. 1, J.A.W. Coetzer & R.C. Tustin, Oxford University Press, Oxford, 486-497.

Pizza, M., Scarlato, V., Masignani, V., Giuliani, M.M., Arico, B., Comanducci, M., Jennings, G.T., Baldi, L., Bartolini, E., Capecchi, B., Galeotti, C.L., Luzzi, E., Manetti, R., Marchetti, E., Mora, M., Nuti, S., Ratti, G., Santini, L., Savino, S., Scarselli, M., Storni, E., Zuo, P., Broeker, M., Hundt, E., Knapp, B., Blair, E., Mason, T., Tettelin, H., Hood, D.W., Jeffries, A.C., Saunders, N.J., Granoff, D.M., Venter, J.C., Moxon, E.R., Grandi, G., & Rappuoli, R. (2000). Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing, *Science*, **287**, 1816-1820.

Plotkin, J.B., Dushoff, J., & Fraser, H.B. (2004). Detecting selection using a single genome sequence of *M. tuberculosis* and *P. falciparum*, *Nature*, **428**, 942-945.

Polley, S.D. & Conway, D.J. (2001). Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene, *Genetics*, **158**, 1505-1512.

Pond, S.L. & Frost, S.D. (2005a). Datamonkey: rapid detection of selective pressure on individual sites of codon alignments, *Bioinformatics.*, **21**, 2531-2533.

Pond, S.L. & Frost, S.D. (2005b). Not so different after all: a comparison of methods for detecting amino acid sites under selection, *Mol. Biol. Evol.*, **22**, 1208-1222.

Pond, S.L., Frost, S.D., & Muse, S.V. (2005). HyPhy: hypothesis testing using phylogenies, *Bioinformatics.*, **21**, 676-679.

Preston, P.M. & Brown, C.G. (1985). Inhibition of lymphocyte invasion by sporozoites and the transformation of trophozoite infected lymphocytes *in vitro* by serum from *Theileria annulata* immune cattle, *Parasite Immunol.*, **7**, 301-314.

Preston, P.M. & Brown, C.G. (1988). Macrophage-mediated cytostasis and lymphocyte cytotoxicity in cattle immunized with *Theileria annulata* sporozoites or macroschizont-infected cell lines, *Parasite Immunol.*, **10**, 631-647.

Preston, P.M., Brown, C.G., Bell-Sakyi, L., Richardson, W., & 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.

Preston, P.M., Brown, C.G., Entrican, G., Richardson, W., & Boid, R. (1993). Synthesis of tumour necrosis factor-alpha and interferons by mononuclear cells from *Theileria annulata*-infected cattle, *Parasite Immunol.*, **15**, 525-534.

Preston, P.M., Brown, C.G., & Spooner, R.L. (1983). Cell-mediated cytotoxicity in *Theileria annulata* infection of cattle with evidence for BoLA restriction, *Clin. Exp. Immunol.*, **53**, 88-100.

Preston, P.M., Hall, F.R., Glass, E.J., Campbell, J.D., Darghouth, M.A., Ahmed, J.S., Shiels, B.R., Spooner, R.L., Jongejan, F., & Brown, C.G. (1999). Innate and adaptive immune responses co-operate to protect cattle against *Theileria annulata*, *Parasitol. Today*, **15**, 268-274.

Ptak, S.E. & Przeworski, M. (2002). Evidence for population growth in humans is confounded by fine-scale population structure, *Trends Genet.*, **18**, 559-563.

Purnell, R.E. (1978). *Theileria annulata* as a hazard to cattle in countries in the northern Mediterranean littoral, *Veterinary Sciences Communications*, **2**, 3-10.

Rajendram, D., Ayenza, R., Holder, F.M., Moran, B., Long, T., & Shah, H.N. (2006). Long-term storage and safe retrieval of DNA from microorganisms for molecular analysis using FTA matrix cards, *J. Microbiol. Methods*.

Rand, D.M. & Kann, L.M. (1996). Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from *Drosophila*, mice, and humans, *Mol. Biol. Evol.*, **13**, 735-748.

Rasti, N., Wahlgren, M., & Chen, Q. (2004). Molecular aspects of malaria pathogenesis, *FEMS Immunol. Med. Microbiol.*, **41**, 9-26.

Razakandrainibe, F.G., Durand, P., Koella, J.C., de Meeus, T., Rousset, F., Ayala, F.J., & Renaud, F. (2005). 'Clonal' population structure of the malaria agent *Plasmodium falciparum* in high-infection regions, *Proc. Natl. Acad. Sci. USA*, **102**, 17388-17393.

Rich, S.M., Hudson, R.R., & Ayala, F.J. (1997). *Plasmodium falciparum* antigenic diversity: evidence of clonal population structure, *Proc. Natl. Acad. Sci. USA*, **94**, 13040-13045.

Richardson, J.O., Forsyth, L.M., Brown, C.G., & Preston, P.M. (1998). Nitric oxide causes the macroschizonts of *Theileria annulata* to disappear and host cells to become apoptotic, *Vet. Res. Commun.*, **22**, 31-45.

Robert, L., Jacob, M.P., Fulop, T., Timar, J., & Hornebeck, W. (1989). Elastonectin and the elastin receptor, *Pathol. Biol. (Paris)*, **37**, 736-741.

Roberts, D.J., Craig, A.G., Berendt, A.R., Pinches, R., Nash, G., Marsh, K., & Newbold, C.I. (1992). Rapid switching to multiple antigenic and adhesive phenotypes in malaria, *Nature*, **357**, 689-692.

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

Rogers, R.J., Dimmock, C.K., de Vos, A.J., & Rodwell, B.J. (1988). Bovine leucosis virus contamination of a vaccine produced *in vivo* against bovine babesiosis and anaplasmosis, *Aust. Vet. J.*, **65**, 285-287.

Ross, B.C., Czajkowski, L., Hocking, D., Margetts, M., Webb, E., Rothel, L., Patterson, M., Agius, C., Camuglia, S., Reynolds, E., Littlejohn, T., Gaeta, B., Ng, A., Kuczek, E.S., Mattick, J.S., Gearing, D., & Barr, I.G. (2001). Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis*, *Vaccine*, **19**, 4135-4142.

Rozas, J. & Rozas, R. (1999). DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis, *Bioinformatics*, **15**, 174-175.

Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X., & Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods, *Bioinformatics*, **19**, 2496-2497.

Sager, H., Bertoni, G., & Jungi, T.W. (1998). Differences between B cell and macrophage transformation by the bovine parasite, *Theileria annulata*: a clonal approach, *J. Immunol.*, **161**, 335-341.

Sager, H., Davis, W.C., Dobbelaere, D.A., & Jungi, T.W. (1997). Macrophage-parasite relationship in theileriosis. Reversible phenotypic and functional dedifferentiation of macrophages infected with *Theileria annulata*, *J. Leukoc. Biol.*, **61**, 459-468.

Saitou, N. & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Mol. Biol. Evol.*, **4**, 406-425.

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

Salih, D.E., Ahmed, J.S., Bakheit, M.A., Ali, E.B., El Hussein, A.M., Hassan, S.M., Shariff, O.E., Fadl, M., & Jongejan, F. (2005). Validation of the indirect *TaSP* enzyme-linked immunosorbent assay for diagnosis of *Theileria annulata* infection in cattle, *Parasitol. Res.*, **97**, 302-308.

Samantaray, S.N., Bhattacharyulu, Y., & 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.

Samish, M. (1977). Infective *Theileria annulata* in the tick without a blood meal stimulus, *Nature*, **270**, 51-52.

Sangwan, A.K., Chhabra, M.B., & Samantaray, S. (1989). Relative role of male and female *Hyalomma anatolicum anatolicum* ticks in *Theileria* transmission, *Vet. Parasitol.*, **31**, 83-87.

Sayin, F., Dincer, S., Karaer, Z., Cakmak, A., Inci, A., Ali Yukari, B., & Eren, H. (1991). Epidemiological study on tropical theileriosis around Ankara (Conference Proceedings), National Dairy Development Board, Anand, India, 51-54.

Sayin, F., Dincer, S., Karaer, Z., Cakmak, A., Inci, A., Yukari, B.A., Eren, H., Vatansever, Z., & Nalbantoglu, S. (2003). Studies on the epidemiology of tropical theileriosis (*Theileria annulata* infection) in cattle in Central Anatolia, Turkey, *Trop. Anim. Health Prod.*, **35**, 521-539.

Schein, E., Buscher, G., & 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], *Z. Parasitenkd.*, **48**, 123-136.

Schein, E. & 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], *Z. Parasitenkd.*, **56**, 287-303.

Schlotterer, C. (2000). Evolutionary dynamics of microsatellite DNA, *Chromosoma*, **109**, 365-371.

Schlotterer, C. (2003). Hitchhiking mapping--functional genomics from the population genetics perspective, *Trends Genet.*, **19**, 32-38.

Schnittger, L., Hong, Y., Jianxun, L., Ludwig, W., Shayan, P., Rahbari, S., Voss-Holtmann, A., & Ahmed, J.S. (2000). Phylogenetic analysis by rRNA comparison of the highly pathogenic sheep-infecting parasites *Theileria lestoquardi* and a *Theileria* species identified in China, *Ann. N. Y. Acad. Sci.*, **916**, 271-275.

Schnittger, L., Katzer, F., Biermann, R., Shayan, P., Boguslawski, K., McKellar, S., Beyer, D., Shiels, B.R., & Ahmed, J.S. (2002). Characterization of a polymorphic *Theileria annulata* surface protein (TaSP) closely related to PIM of *Theileria parva*: implications for use in diagnostic tests and subunit vaccines, *Mol. Biochem. Parasitol.*, **120**, 247-256.

Schnittger, L., Yin, H., Gubbels, M.J., Beyer, D., Niemann, S., Jongejan, F., & Ahmed, J.S. (2003). Phylogeny of sheep and goat *Theileria* and *Babesia* parasites, *Parasitol. Res.*, **91**, 398-406.

Schreuder, B.E., Uilenberg, G., & Tondeur, W. (1977). Studies on Theileriidae (Sporozoa) in Tanzania. VIII. Experiments with African buffalo (*Syncerus caffer*), *Tropenmed. Parasitol.*, **28**, 367-371.

Sergent, E., Donatien, A., Parrot, L., & Lestoquard, F. (1945). *Etudes sur les piroplasmoses bovines,* Institut Pastuer d'Algerie, Alger.

Service, M.W. (1997). Mosquito (Diptera: Culicidae) dispersal--the long and short of it, *J. Med. Entomol.*, **34**, 579-588.

Shapiro, S.Z., Fujisaki, K., Morzaria, S.P., Webster, P., Fujinaga, T., Spooner, P.R., & Irvin, A.D. (1987). A life-cycle stage-specific antigen of *Theileria parva* recognized by anti-macroschizont monoclonal antibodies, *Parasitology*, **94 (Pt 1)**, 29-37.

Sharp, P.M. (2005). Gene "volatility" is most unlikely to reveal adaptation, *Mol. Biol. Evol.*, **22**, 807-809.

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

Sharpe, R.T. & Langley, A.M. (1983). The effect of *Theileria annulata* infection on the immune response of cattle to foot-and-mouth disease, *Br. Vet. J.*, **139**, 378-385.

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), *J. Cell Sci.*, **101 (Pt 1)**, 109-123.

Shiels, B., Hall, R., Glascodine, J., McDougall, C., Harrison, C., Taracha, E., Brown, D., & Tait, A. (1989). Characterization of surface polypeptides on different life-cycle stages of *Theileria annulata*, *Mol. Biochem. Parasitol.*, **34**, 209-220.

Shiels, B., Kinnaird, J., McKellar, S., Dickson, J., Miled, L.B., Melrose, R., Brown, D., & Tait, A. (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 (Pt 1)**, 99-107.

Shiels, B., McDougall, C., Tait, A., & Brown, C.G. (1986). Antigenic diversity of *Theileria annulata* macroschizonts, *Vet. Parasitol.*, **21**, 1-10.

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**, 1659-1670.

Shiels, B.R. (1999). Should I stay or should I go now? A stochastic model of stage differentiation in *Theileria annulata*, *Parasitol. Today*, **15**, 241-245.

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

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

Simonsen, K.L., Churchill, G.A., & Aquadro, C.F. (1995). Properties of statistical tests of neutrality for DNA polymorphism data, *Genetics*, **141**, 413-429.

Singh, D.K., Jagdish, S., Gautam, O.P., & 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, D.K., Thakur, M., Raghav, P.R., & Varshney, B.C. (1993). Chemotherapeutic trials with four drugs in crossbred calves experimentally infected with *Theileria annulata*, *Res. Vet. Sci.*, **54**, 68-71.

Skilton, R.A., Bishop, R.P., Wells, C.W., Spooner, P.R., Gobright, E., Nkonge, C., Musoke, A.J., Macklin, M., & Iams, K.P. (1998). Cloning and characterization of a 150 kDa microsphere antigen of *Theileria parva* that is immunologically cross-reactive with the polymorphic immunodominant molecule (*PIM*), *Parasitology*, **117 (Pt 4)**, 321-330.

Skilton, R.A., Musoke, A.J., Nene, V., Wasawo, D.P., Wells, C.W., Spooner, P.R., Bishop, R.P., Osaso, J., Nkonge, C., Latif, A., & Morzaria, S.P. (2000). Molecular characterisation of a *Theileria lestoquardi* gene encoding a candidate sporozoite vaccine antigen, *Mol. Biochem. Parasitol.*, **107**, 309-314.

Smith, J.M., Smith, N.H., O'Rourke, M., & Spratt, B.G. (1993). How clonal are bacteria?, *Proc. Natl. Acad. Sci. USA*, **90**, 4384-4388.

Smith, J.R., Carpten, J.D., Brownstein, M.J., Ghosh, S., Magnuson, V.L., Gilbert, D.A., Trent, J.M., & Collins, F.S. (1995). Approach to genotyping errors caused by nontemplated nucleotide addition by *Taq* DNA polymerase, *Genome Res.*, **5**, 312-317.

Somerville, R.P., Littlebury, P., Pipano, E., Brown, C.G., Shkap, V., Adamson, R.E., Oliver, R.A., Glass, E.J., & Hall, F.R. (1998). Phenotypic and genotypic alterations associated with the attenuation of a *Theileria annulata* vaccine cell line from Turkey, *Vaccine*, **16**, 569-575.

Spitalska, E., Torina, A., Cannella, V., Caracappa, S., & Sparagano, O.A. (2004). Discrimination between *Theileria lestoquardi* and *Theileria annulata* in their vectors and hosts by RFLP based on the 18S rRNA gene, *Parasitol. Res.*, **94**, 318-320.

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

Stagg, D.A., Bishop, R.P., Morzaria, S.P., Shaw, M.K., Wesonga, D., Orinda, G.O., Grootenhuis, J.G., Molyneux, D.H., & Young, A.S. (1994). Characterization of *Theileria parva* which infects waterbuck (*Kobus defassa*), *Parasitology*, **108 (Pt 5)**, 543-554.

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)**, 341-346.

Sutherland, I.A., Shiels, B.R., Jackson, L., Brown, D.J., Brown, C.G., & Preston, P.M. (1996). *Theileria annulata*: altered gene expression and clonal selection during continuous *in vitro* culture, *Exp. Parasitol.*, **83**, 125-133.

Suzuki, Y. & Gojobori, T. (1999). A method for detecting positive selection at single amino acid sites, *Mol. Biol. Evol.*, **16**, 1315-1328.

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

Swan, D.G., Stadler, L., Okan, E., Hoffs, M., Katzer, F., Kinnaird, J., McKellar, S., & Shiels, B.R. (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.

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

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism, *Genetics*, **123**, 585-595.

Takezaki, N. & Nei, M. (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA, *Genetics*, **144**, 389-399.

Talman, A.M., Domarle, O., McKenzie, F.E., Ariey, F., & Robert, V. (2004). Gametocytogenesis: the puberty of *Plasmodium falciparum*, *Malar. J.*, **3**, 24.

Tanabe, K., Sakihama, N., Farnert, A., Rooth, I., Bjorkman, A., Walliker, D., & Ranford-Cartwright, L. (2002). *In vitro* recombination during PCR of *Plasmodium falciparum* DNA: a potential pitfall in molecular population genetic analysis, *Mol. Biochem. Parasitol.*, **122**, 211-216.

Tanaka, M., Ohgitani, T., Okabe, T., Kawamoto, S., Takahashi, K., Onuma, M., Kawakami, Y., & Sasaki, N. (1990). Protective effect against intraerythrocytic merozoites of *Theileria sergenti* infection in calves by passive transfer of monoclonal antibody, *Nippon Juigaku. Zasshi*, **52**, 631-633.

Taracha, E.L., Goddeeris, B.M., Teale, A.J., Kemp, S.J., & Morrison, W.I. (1995). Parasite strain specificity of bovine cytotoxic T cell responses to *Theileria parva* is determined primarily by immunodominance, *J. Immunol.*, **155**, 4854-4860.

Tautz, D. & Renz, M. (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes, *Nucleic Acids Res.*, **12**, 4127-4138.

Taylor, L.H., Katzer, F., Shiels, B.R., & Welburn, S.C. (2003). Genetic and phenotypic analysis of Tunisian *Theileria annulata* clones, *Parasitology*, **126**, 241-252.

Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C., Nelson, K.E., Eisen, J.A., Ketchum, K.A., Hood, D.W., Peden, J.F., Dodson, R.J., Nelson, W.C., Gwinn, M.L., DeBoy, R., Peterson, J.D., Hickey, E.K., Haft, D.H., Salzberg, S.L., White, O., Fleischmann, R.D., Dougherty, B.A., Mason, T., Ciecko, A., Parksey, D.S., Blair, E., Cittone, H., Clark, E.B., Cotton, M.D., Utterback, T.R., Khouri, H., Qin, H., Vamathevan, J., Gill, J., Scarlato, V., Masignani, V., Pizza, M., Grandi, G., Sun, L., Smith, H.O., Fraser, C.M., Moxon, E.R., Rappuoli, R., & Venter, J.C. (2000). Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58, *Science*, **287**, 1809-1815.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., & Higgins, D.G. (1997). The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools, *Nucleic Acids Res.*, **25**, 4876-4882.

Tibayrenc, M., Kjellberg, F., & Ayala, F.J. (1990). A clonal theory of parasitic protozoa: the population structures of *Entamoeba*, *Giardia*, *Leishmania*, *Naegleria*, *Plasmodium*, *Trichomonas*, and *Trypanosoma* and their medical and taxonomical consequences, *Proc. Natl. Acad. Sci. USA*, **87**, 2414-2418.

Tisdell, C.A., Harrison, S.R., & Ramsay, G.C. (1999). The economic impacts of endemic diseases and disease control programmes, *Rev. Sci. Tech.*, **18**, 380-398.

Toye, P., Gobright, E., Nyanjui, J., Nene, V., & Bishop, R. (1995a). Structure and sequence variation of the genes encoding the polymorphic, immunodominant molecule (*PIM*), an antigen of *Theileria parva* recognized by inhibitory monoclonal antibodies, *Mol. Biochem. Parasitol.*, **73**, 165-177.

Toye, P., Nyanjui, J., Goddeeris, B., & Musoke, A.J. (1996). Identification of neutralization and diagnostic epitopes on PIM, the polymorphic immunodominant molecule of *Theileria parva*, *Infect. Immun.*, **64**, 1832-1838.

Toye, P.G., Goddeeris, B.M., Iams, K., Musoke, A.J., & Morrison, W.I. (1991). Characterization of a polymorphic immunodominant molecule in sporozoites and schizonts of *Theileria parva*, *Parasite Immunol.*, **13**, 49-62.

Toye, P.G., Metzelaar, M.J., Wijngaard, P.L., Nene, V., Iams, K., Roose, J., Nyanjui, J.K., Gobright, E., Musoke, A.J., & Clevers, H.C. (1995b). Characterization of the gene encoding the polymorphic immunodominant molecule, a neutralizing antigen of *Theileria parva*, *J. Immunol.*, **155**, 1370-1381.

Tsur, I. & Pipano, E. (1966). Attenuation of virulence of strains of *Theileria annulata* by growth and passage through tissue culture, *J. Protozool.*, **13**, 33-34.

Turner, A.J. (1994). PIG-tailed membrane proteins, *Essays Biochem.*, **28**, 113-127.

Uilenberg, G. (1981). Theilerial species of domestic livestock, in *Advances in the Control of Theileriosis*, A.D. Irvin, M.P. Cunningham, & A.S. Young, Martinus Nijhoff, The Hague, 4-37.

Urdaneta, L., Lal, A., Barnabe, C., Oury, B., Goldman, I., Ayala, F.J., & Tibayrenc, M. (2001). Evidence for clonal propagation in natural isolates of *Plasmodium falciparum* from Venezuela, *Proc. Natl. Acad. Sci. USA*, **98**, 6725-6729.

Verra, F. & Hughes, A.L. (1999). Natural selection on apical membrane antigen-1 of *Plasmodium falciparum*, *Parassitologia*, **41**, 93-95.

Visser, A.E., Abraham, A., Sakyi, L.J., Brown, C.G., & Preston, P.M. (1995). Nitric oxide inhibits establishment of macroschizont-infected cell lines and is produced by macrophages of calves undergoing bovine tropical theileriosis or East Coast fever, *Parasite Immunol.*, **17**, 91-102.

von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites, *Nucleic Acids Res.*, **14**, 4683-4690.

Walker, A.R., McKellar, S.B., Bell, L.J., & Brown, C.G. (1979). Rapid quantitative assessment of *Theileria* infection in ticks, *Trop. Anim. Health Prod.*, **11**, 21-26.

Walker-Jonah, A., Dolan, S.A., Gwadz, R.W., Panton, L.J., & Wellems, T.E. (1992). An RFLP map of the *Plasmodium falciparum* genome, recombination rates and favored linkage groups in a genetic cross, *Mol. Biochem. Parasitol.*, **51**, 313-320.

Wayne, M.L. & Simonsen, K.L. (1998). Statistical tests of neutrality in the age of weak selection, *Trends in Ecology and Evolution*, **13**, 236-240.

Weir, B.S. & Cockerham, C.C. (1984). Estimating F-statistics for the analyses of population structure, *Evolution*, **38**, 1358-1370.

Wilkie, G.M., Brown, C.G., Kirvar, B.E., Thomas, M., Williamson, S.M., Bell-Sakyi, L.J., & Sparagano, O. (1998). Chemoprophylaxis of *Theileria annulata* and *Theileria parva* infections of calves with buparvaquone, *Vet. Parasitol.*, **78**, 1-12.

Williamson, S. (1991). Studies on Immunization using inactivated sporozoites of *Theileria annulata* (Conference Proceedings), National Dairy Development Board, Anand, India, 99-100.

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

Wizemann, T.M., Heinrichs, J.H., Adamou, J.E., Erwin, A.L., Kunsch, C., Choi, G.H., Barash, S.C., Rosen, C.A., Masure, H.R., Tuomanen, E., Gayle, A., Brewah, Y.A., Walsh, W., Barren, P., Lathigra, R., Hanson, M., Langermann, S., Johnson, S., & Koenig, S. (2001). Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection, *Infect. Immun.*, **69**, 1593-1598.

Wootton, J.C., Feng, X., Ferdig, M.T., Cooper, R.A., Mu, J., Baruch, D.I., Magill, A.J., & Su, X.Z. (2002). Genetic diversity and chloroquine selective sweeps in *Plasmodium falciparum*, *Nature*, **418**, 320-323.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood, *Comput. Appl. Biosci.*, **13**, 555-556.

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.

Zablotskii, V. T. (1991). Specific prevention of bovine theileriosis in the Soviet Union (Conference Proceedings), National Dairy Development Board, Anand, India, 9-10.

Zhang, G.L., Khan, A.M., Srinivasan, K.N., August, J.T., & Brusic, V. (2005). MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides, *Nucleic Acids Res.*, **33**, W172-W179.

Zhang, J. (2005). On the evolution of codon volatility, *Genetics*, **169**, 495-501.

Zhang, Z. H. (1991). *Theileria annulata* and it's control in China (Conference Proceedings), National Dairy Development Board, Anand, India, 11-14.

Zhivotovsky, L.A. & Feldman, M.W. (1995). Microsatellite variability and genetic distances, *Proc. Natl. Acad. Sci. USA*, **92**, 11549-11552.

Zhu, J. & Hollingdale, M.R. (1991). Structure of *Plasmodium falciparum* liver stage antigen-1, *Mol. Biochem. Parasitol.*, **48**, 223-226.

Zhuang, W.Z., Sugimoto, C., Kubota, S., Onoe, S., & Onuma, M. (1995). Antigenic alteration in major piroplasm surface proteins of *Theileria sergenti* during infection, *Vet. Parasitol.*, **60**, 191-198.