COMPARATIVE ANNOTATION AND ANALYSIS OF PROTEIN-CODING DNA SEQUENCES (CDS) OF *THEILERIA PARVA* MARIKEBUNI AND *THEILERIA PARVA* MUGUGA GENOMES

BY

OBIERO, GEORGE FREDRICK OPONDO

A THESIS SUBMITTED IN PARTIAL FULFILLMENT FOR THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN CELL AND MOLECULAR BIOLOGY.

DEPARTMENT OF ZOOLOGY

MASENO UNIVERSITY

ABSTRACT

Theileria parva Muguga genome was the first to be sequenced and published in this genus of great economic, veterinary and biological importance to livestock industry in East and Central Africa. The aim was to aid in identification of schizont antigens for vaccine development and to enhance comparative genomics with other related apicomplexas. To add to the repertoire of resource base, *T. parva* Marikebuni was recently sequenced because of its high genotypic diversity given that its infection cannot be cross-protected by the Muguga cocktail vaccine. Reported here is the annotation and curation of protein-coding DNA sequences (CDS) of the partial genome of *T. parva* Marikebuni against *T. parva* Muguga via Artemis Comparison Tool (ACT). The genome was analyzed for both strain-specific micro- and macro-satellite markers (also called variable number tandem repeats, VNTRs) and codon usage bias of coding open reading frames (ORFs) containing the VNTR markers. The results reported here showed that *T. parva* Marikebuni has a compact but protracted nuclear genome, encoding over 3900 CDS. The majority of these CDS are predicted as multi-exonic, but with lower count number relative to those observed in *T. parva* Muguga genome. The genome is AT-rich with about 32.64% GC content and shares a perfect synteny with the template genome in terms of gene structure and nucleotide composition. The CDS were assigned unique feature identifiers including putative functions from database gene ontologies (GO). This study was able to characterize and locate the VNTRs within both genomes. Most VNTRs were found to be located in the non-coding regions of the genomes but peculiarly, seven of them were located in exonic ORFs. The codon usages in these genomes are biased towards AT-rich codons as would positively be expected of AT-rich genomes. Statistical analysis at 0.05 confidence showed that there was no significant difference in the codon usage both within and between the *Theileria* genomes. These results will be crucial in building of *T. parva* database and make mining of the micro- and macro-satellites for any future comparative studies achievable. The findings will further enhance the search and prediction of the many *T. parva* genes with unknown functions. In addition, the present study will aid in definition of strain-specific markers if the whole *T. parva* Marikebuni genome can be completely sequenced.

CHAPTER ONE

INTRODUCTION

1.1. Background information

Theileria parva is an intracellular hemo-protozoan parasite transmitted by the brown ear tick, *Rhipicephalus appendiculatus*. The parasite causes the devastating lympho-proliferative East Coast fever (ECF) in cattle in East and Central Africa, with very heavy economic burden (Norval *et al.*, 1992). The principal agents are *T. parva* Muguga and *T. parva* Marikebuni. The other *T. parva* strains and field isolates include Serengeti, Mariakani, Boleni, Zimbabwe, Kakuzi, Buffalo, Uganda, Kiambu and Zambia, among others. *Theileria* parasites have been of considerable biological interest, as they are the only eukaryotic pathogens known to reversibly transform lymphocytes (Dobbelaere and McKeever, 2002); however, the genomic complexities of the species and effective control of ECF are yet to be resolved. The mechanisms by which *Theileria* sporozoites invade lymphocytes, escape from the invasion vacuole, interact with the host cell cytoskeleton, and alter cellular signaling pathways are incompletely understood (Roos, 2005). Previous studies showed that *T. parva* is both phenotypically and genotypically heterogeneous, revealing also that multi-locus genotypes decrease with each passage in the tick vector (Goddeeris *et al.*, 1990). In an attempt to explain this observed heterogeneity, other investigators, through clonal genotyping in *T. parva* Marikebuni, indicated that meiotic genotypic recombination occurs among various species, giving rise to shuffling of immunological determinants, thus allowing the immunologic evasion at herd level (Katzer *et al.*, 2006). Much of this diversity is known to arise in the gut of tick vectors and may later change especially in salivary type III acinar cells (Katzer *et al.*, 2006). Even though physical cross-overs have been reported to occur between polymorphic variable number tandem repeats (VNTR) loci (two on chromosome 1, two on chromosome 2, three on chromosome 3, and one on chromosome 4) (Westesson and Holmes, 2009), the ratio of point mutation to recombination and their extent across the genome in terms of allelic polymorphism is poorly known.

Previous studies reported complete genome sequences for *T. parva* Muguga and *T.*

annulata, which were estimated to be approximately 8.3 and 8.5 Mbp respectively, with four chromosomes each (Gardner *et al.*, 2005; Pain *et al.*, 2005). The chromosomes were reported to be extremely AT-rich, especially at putative centromeric regions, as is the case in *P. falciparum* (Gardner *et al.*, 2005), but has much less complex sub-telomeric regions than the latter. Previous analysis of the *T. parva* genome showed that 61% of predicted genes have no significant similarity to previously known sequences, even to proteins of related protozoa such as *Plasmodium* and, therefore, have no predicted function (Gardner *et al.*, 2005). Automated annotation, using a combination of two Markov chain model-based gene prediction programs, GlimmerM and PHAT (phase transform) demonstrated that *T. parva* nuclear genomes encode 4036 protein coding genes (Shah *et al.*, 2006). This was 20% fewer than *P. falciparum*, but exhibited higher gene density, a greater proportion of genes with introns, and shorter intergenic regions (IGRs) (Gardner *et al.*, 2005). Other studies reported that a large fraction of *Theileria* non-coding DNA is kept constant by purifying selection (Guo and Silva, 2008). This high conservation rate confirms the functional importance of non-coding sequences in *Theileria*, which goes beyond a role of passive intergenic spacers. This assertion is further supported by the higher degree of sequence conservation in IGRs that border the 5' end of genes relative to what is observed in IGRs flanked by termination codons, since IGR sequence conservation between species in regions upstream of genes, is associated with the presence of regulatory elements (Guo and Silva, 2008). However, the non-coding regions of *T. parva* genomes remain remarkably unresolved, and little is known about the forces that shape them. As a whole, the genomic structure and statistics of *T. parva* Marikebuni need to be established.

Like many other parasitic protozoa, *Theileria* spp. have been reported to have tandem arrays of genus-specific, hyper-variable gene families that map adjacent to the telomeres with an overall arrangement that appears conserved (Bishop *et al.*, 2000; Roos, 2005). The numerical advantage of this telomeric location is perhaps to allow rapid evolution of surface or secreted proteins, which are crucial in the survival of these intra-cytoplasmic parasites. The other areas with tandem arrays include the most rapidly evolving and dispersed multi-copy protein family of *Tpr* (*T. parva* repeats) whose assembly has been known to be difficult because of their array lengths, and an abundant sporozoite surface antigen, *p67*, a primary

target of parasite neutralizing antibodies (Pain *et al.*, 2005; Bishop *et al.*, 2009). The locus-specificity of multi-copy telomere associated genes and centromeric repeats, the extent of dispersion and polymorphism of the *Tpr* gene family and *p67* antigen remains to be confirmed in the *T. parva* Marikebuni strain.

An earlier study described the cloning and characterization of two 80% AT-rich mini-satellites in *T. parva* Muguga and *T. parva* Uganda isolates (Bishop *et al.*, 1998). Unusually, the sequences showed interstitial-even distribution across the genome and the tandem repeats were reported to be interrupted by other short sequences (Thompson *et al.*, 2001). The observed size polymorphism at the loci studied was attributed to the variable copy number of the mini-satellite tandem repeats, supported by later observations made in *Plasmodium* species (Thompson *et al.*, 2001). A recent study comprehensively revealed extensive genetic diversity at VNTR loci and existence of linkage disequilibrium among different field isolates of *T. parva* from Kenya (Odongo *et al.*, 2006). The results contrasted previous reports on lower vertebrates whose mini-satellites are GC-rich and are biased to sub-telomeric regions, with few exceptions in humans, where the AT-rich sequences tend to be locus-specific (Vogt, 1990; Oura *et al.*, 2003; Weir *et al.*, 2007). One of the studies reported these polymorphic molecular markers to be widely distributed across the *T. parva* genome, thus setting a step towards both population and comparative studies (Oura *et al.*, 2003). The mechanism of mutation of micro-satellites and mini-satellites in non-coding regions is different, but their frequency is typically higher than that of base substitutions in open reading frames (ORFs) encoding proteins. The high rates of mutation result in high levels of polymorphism at the VNTR loci, making them ideally suited for high resolution molecular fingerprinting of pathogen isolates, forensic science, evolutionary biology and genomic analysis (Burke, 1991; Bishop *et al.*, 1998). For these reasons, the VNTRs have been used for quality control of *T. parva* stabilates and in monitoring deployment of infection-and-treatment of live vaccination in the field (Bishop *et al.*, 2009). Most previous molecular epidemiology researches on *T. parva* have mainly been based on restriction fragment length polymorphism (RFLP) combined with PCR to analyze the multi-copy gene families and mini-satellite loci (Pain *et al.*, 2005). The present study used molecular marker primer pairs rather than RFLP (which is limited to identifying only few polymorphisms), to

comparatively study genotypic polymorphism surrounding VNTRs genomes of *T. parva* Muguga and *T. parva* Marikebuni.

1.2. Statement of the study problem

The remedy for the control of *T. parva*, which reversibly transforms the bovine host lymphocytes by manipulating and synchronizing its life-cycle to that of the host cell, still remains elusive. Evolutionary studies on all members of the *Apicomplexa* have not been feasible due to lack of genomic information, especially in this important genus of *Theileria*. No comparative genomic study has been reported on the *T. parva* Marikebuni genome, which was previously sequenced using whole genome shot-gun technique by the Bioinformatics group at International Livestock Research Institute (ILRI). The genomic variation of variable number tandem repeat (VNTR) sequences between *T. parva* Marikebuni and *T. parva* Muguga is unknown. In addition, the precise loci of known molecular markers in the Marikebuni genome are yet to be resolved. This study, as the first comparative genomic study of related strains of *T. parva*, annotated and curated the protein-coding DNA sequences (CDS) and characterized variable number tandem repeats (VNTRs) loci in the genomes through *in silico* approaches with confirmatory *in vitro* PCR.

1.3. Justification and significance of the study

The genomic studies on *T. parva* have already shown remarkable differences from the other apicomplexan genomes sequenced to date and there is need to increase the number of annotated geneomes of *Theileria* spp. Vaccinations using *T. parva* Marikebuni stabilate have been shown to protect against heterologous challenges including that of *T. parva* Muguga, yet the genotypic diversity of the *T. parva* Marikebuni genome is unknown. The outcome of this study has added value to the understanding of the genotypic diversities of *T. parva,* and possibly will improve search for intra-species-specific molecular markers that could potentially be used in future studies to amplify genomes from *T. parva* field isolates. The availability of more comparatively analyzed *T. parva* genome sequences, will significantly add power to the search and revelation of putative functions of conserved regions that will advance, in general, the study of gene regulation in *Theileria* as

hypothesized previously (Behnke *et al.*, 2008; Sunil *et al.*, 2008).

1.4. Research questions

a) How do the protein-coding sequences (CDS) compare between *T. parva* Marikebuni and *T. parva* Muguga genomes?

b) What are the characteristics of variable number tandem repeats (VNTRs) in the genomes of *T. parva* Marikebuni and *T. parva* Muguga?

c) How does the presence of variable number tandem repeats (VNTRs) affect codon usage in the genomic open reading frames (ORFs) of *T. parva* Marikebuni as compared to *T. parva* Muguga?

1.5. Study Objective

1.5.1. Main objective

To comparatively annotate and curate the protein-coding sequences (CDS) and to characterize the variable number tandem repeats (VNTRs) loci in open reading frames (ORFs) between *T. parva* Marikebuni and *T. parva* Muguga strains.

1.5.2. Specific objectives

a) To determine the genomic location of protein-coding sequences (CDS) in *T. parva* Marikebuni.

b) To determine the characterizations of the variable number tandem repeats (VNTRs) in the genomes of *T. parva* Marikebuni and *T. parva* Muguga.

c) To determine the codon usage in the genomic open reading frames (ORFs) containing variable number tandem repeats (VNTRs) in both *T. parva* Marikebuni and *T. parva* Muguga genomes.

CHAPTER TWO
LITERATURE REVIEW

2.0 Introduction

The high mortality and morbidity in cattle associated with East Coast fever (ECF), the lympho-proliferative disease caused by *T. parva,* are leading causes of heavy socio-economic losses to, especially, small-holder farmers in sub-Saharan, central and eastern Africa (Norval *et al.*, 1992). Characteristically, the apicomplexan protozoan life cycles have an expansive asexual reproductive phase (Shirley and Harvey, 2000). *T. parva* has an obligatory short sexual cycle with a transient diploid stint, a pre-requisite for genetic recombination and random assortment of alleles in the vector tick gut. This relates directly to the high endemicity observed in East Africa where both bovine and tick-vector hosts get infected with multiple isolates having varied genotypes (Oura *et al.*, 2003). The complex life-cycle stages of *Theileria* (**Figure 1**), like those of other apicomplexan parasites, are characterized by persistent themes, with subtle variations (Dobbelaere and Kuenzi, 2004). *Theileria* sporozoites and merozoites are unusual in being non-motile. They invade host cells in an irreversibly active manner (Jura *et al.*, 1983; Jura, 1984; Shaw *et al.*, 1991; Shaw, 2003). While *Theileria annulata* sporozoites were shown to consistently attach to, and invade, target bovine lymphoid cells with their basal end, thus incriminating receptor-ligand interactions (Jura *et al.*, 1983; Jura, 1984), while *T. parva* sporozoites were reported to invade host cells in a non-specific manner without any further re-orientation on first contact (Shaw *et al.*, 1991; Shaw, 2003). Members of the genus have rhoptries, the secretory organelles that are part of the apical complex, also suspected to function in modifying the parasite's intracellular parasitophorous vacuole for survival; and the secretion of rhoptries coincides with rupture of the invasion compartment, releasing parasites into the host cell cytoplasm (Jura *et al.*, 1983; Roos, 2005). Another unique aspect of *T. parva* biology is that the biased infection of T- and B-lymphocytes results in a reversibly transformed phenotype of uncontrolled proliferation of the host cells that remain persistently infected, the culprit stages being the intra-lymphocytic schizont stage and the intra-erythrocytic piroplasm stage (Dobbelaere and Kuenzi, 2004). A previous study noted that the complexity of the *T. parva*

life cycle is not matched by a large number of recognizable cell cycle regulators (Gardner *et al.*, 2005), suggestive of the fact that a number of novel regulatory features are yet to be discovered. The host cell microtubules that decorate the surface of schizonts are captured by the host cell spindle during mitosis, favouring infection of both daughter cells (Norval *et al.*, 1992). Some parasite proteins that are thought to modulate host cell phenotype are described in a previous study (Pain *et al.*, 2005). The *T. parva* genome encodes putative secreted forms of EMAP115- and Tau-like proteins, which are missing in *P. falciparum*, but their homologues in higher eukaryotes, are known to interact with microtubules (Gardner *et al.*, 2005). The parasite may also modulate host cell mitosis by influencing disassembly of the host cell spindle via a secreted cdc48-like AAA–adenosine triphosphatase (ATPase associated with diverse cellular activities) (Gardner *et al.*, 2005).

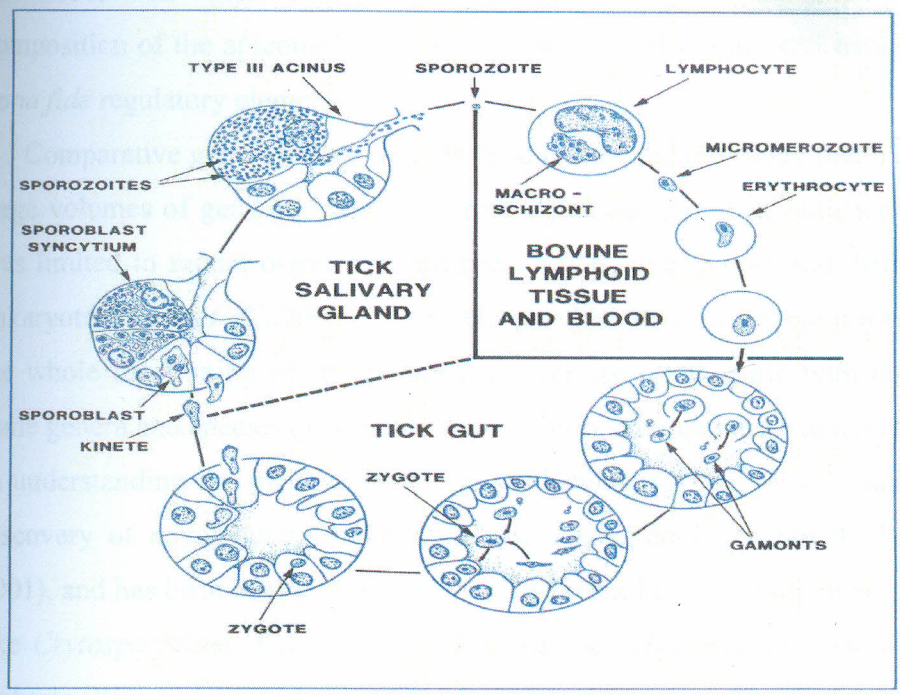

Figure 1. Typical life cycle of *Theileria parva*. The Sporozoites are transmitted to the bovine host in tick saliva. They enter the lymphocytes and develop into macroschizonts. After a period of nuclear division, these give rise to numerous micro-merozoites which escape from the lymphocytes and enter the erythrocytes where they are ingested by feeding ticks. In the tick gut, the parasites differentiate into male and female gamonts which fuse to form zygotes. These enter the cell lining of the gut where they differentiate into kinetes. The kinetes traverse the gut wall and become free in the tick's body cavity. They move to the salivary gland and enter one cell type, the E-cell of the type III acinus. Here, they form an elaborate intracellular sporoblast syncytium which undergoes segmental fission and gives rise to 30,000 to 50,000 sporozoites. These are introduced, with the saliva, into a new mammalian host, initiating a new cycle of parasite development (Shiels *et al.*, 2006).

2.1. Apicomplexan comparative genomic studies

Apicomplexans are generally thought to have reduced transcriptional machinery relative to multi-cellular eukaryotes, with classical promoter elements such as the TATA-box and the CAAT-box reportedly absent in *T. gondii* and *P. falciparum* (Militello *et al.*, 2004). Overall proteomic analysis suggests presence of novel regulatory processes in *Theileria* with a TATA-box binding protein (PF00352) and a transcription binding factor II, TFIIB (PF00382) already reported (Bishop *et al.*, 2009). Recent studies that took into account the genome composition and/or the timing of gene expression reported novel conserved sequence motifs, leading to a considerable expansion of the repertoire of known and putative transcription regulators in this phylum (Behnke *et al.*, 2008; Guo and Silva, 2008; Sunil *et al.*, 2008; van Noort and Huynen, 2006). However, the highly biased nucleotide composition of the apicomplexan genomes sequenced to date still hampers the detection of *bona fide* regulatory elements.

Comparative genomics is a derived field in molecular biology that has enabled mining of large volumes of gene and whole genome databases possible, but its power, until recently, was limited to model organisms and prokaryotes, due to cost and difficulty in sequencing eukaryotes (Pain *et al.*, 2005). The field provides much valuable information in respect to the whole functioning of an organism, as well as, relationship with other members of the same genera and species (Thompson *et al.*, 2001). Comparative analyses have been valuable in understanding the genome evolution, molecular markers for evaluation of vaccines and discovery of novel putative genes, genomic and population structures (Thompson *et al.*, 2001), and has been useful in the study of pathogen-host interactions of protozoan pathogens like *Crytosporidium, Plasmodium, Trypanosoma, Theileria, Leishmania, Toxoplasma,* and *Babesia* (Wilkowsky *et al.*, 2009).

All eukaryotic genomes are composed of gene coding sequences and non-coding sequences [including intergenic regions (IGRs) and introns], and in order to study the evolution and conservation of any genomic region, homologous sequence segments between species or strains are required (Guo and Silva, 2008). However, the exact location of IGRs, introns and exons in most organisms including *T. parva* is often uncertain due to the probabilistic nature of the gene models generated through automated annotation. Non-

coding genome DNA in majority of bacteria and archael genomes is between 6% and 14% but close to 90% in multicellular eukaryotes, with *Theileria* having upto 30% (Guo and Silva, 2008). Unicellular eukaryotes are report to have a higher proportion of non-coding DNA than prokaryotes, but a much more compact genome than multi-cellular eukaryotes (Guo and Silva, 2008). The lengths and numbers of introns vary among taxa, and dramatic differences can be seen across related species. They also differ in their functional non-coding DNA especially telomeric introns, which tend to be longer and more conserved near the 5' than the 3' chromosomal ends and also than those of higher ordinal order (Guo and Silva, 2008). Only a few eukaryotic unicellular parasites have so far been shown to contain transposable elements (Souza *et al.*, 2007), negating their potentiality of being used to explain the differences. Transposable elements, the genetic fragments that often break, transfer and integrate via genetic recombination into other positions and/or a new host genome, are known to commonly occur in prokaryotes and lower parasitic eukaryotes (Chan *et al.*, 2009). However, *Theileria* are known to lack these transposable elements all together (Carlton *et al.*, 2002; Gardner *et al.*, 2005), which probably explains much of the observed difference in the amount of non-coding DNA between these organisms and other eukaryotes. In *T. parva*, it is not clearly known if the reported average IGR distance of 405 bp is sufficient for a complete binding of transcriptones - for example, the *p67* antigen and a downstream ORF having only a 93 bp gap suggests they could be jointly transcribed (Bishop *et al.*, 2009). It is equally unclear whether the genomic non-coding sequences in whole or in-part could explain the observed pathogenicity in the *T. parva* Marikebuni strain, after some conserved regions are reportedly absent in the genome.

2.2 Whole genome sequencing and annotation

The current multiple advances in DNA sequencing techniques have tremendously revolutionized biological genomic researches. These include the whole genome shot-gun (WGS) technique that shears a large genome (in billion base pairs) into smaller fragments that can easily be read separately, then assembled by equally robust computer programs called genome assemblers (Pop, 2009). By use of the vast array of bioinformatics tools, it has become much faster and cheaper to sequence, assemble, annotate and compare the

genomes of both distantly and closely related organisms. However, the presence of repetitive sequences extending thousands to millions of base pairs limits both the sequencing and assembly processes, thus leading to occurrences of gaps and assemblies of contiguous chromosomal sequences (Pop, 2009). The comparative analysis tools have equally unprecedentedly advanced to the present robust ones like the BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi), FASTA (http://www.ebi.ac.uk/Tools/fasta33/index.html), MUMmer (Kurtz *et al.*, 2004), EMBOSS (http://www.ebi.ac.uk), and Artemis Comparative Tool (ACT) (Carver *et al.*, 2008). Initially, it was only possible to do a single gene analysis with either its DNA or amino acid sequences. At present, coupled with the increase in nucleotide and amino acid databases, whole genome comparisons has become handy especially in discovering the evolutionary relationships and functions of the thousands of proteins from hundreds of different organisms (Kurtz *et al.*, 2004).

Currently, *Theileria* whole genome sequences have been reported for *T. parva* Muguga strain (Gardner *et al.*, 2005) against which the *T. parva* Marikebuni strain is being annotated and analyzed, and *T. annulata* whose comparison with the Muguga is already reported in a previous study (Pain *et al.*, 2005). At the time of submission of this work, the other whole genome annotation work is on-going for *T. orientalis* at the Graduate School of Veterinary Medicine, Hokkaido University, Japan (Bishop *et al.*, 2009) and *T. parva* Serengeti, *T. parva* Katete, and *T. parva* Muguga-Marikebuni recombinant at the ILRI Bioinformatics Group lab (un-published). Presently, the recent shot-gun sequenced and comparatively assembled *T. parva* Marikebuni genome, has shown lots of sequence gaps across all its four chromosomes, and has two contigs each in chromosomes 3 and 4. Preliminary annotation and analysis of the genome has been observed to suffer from these gaps, suggesting a need for either a re-sequencing or a re-assembly of the genome to attempt to fill-up the gaps (personal communication).

2.3 Tandem repeat sequences

Tandem repeats in DNA are two or more contiguous, approximate copies of a pattern of nucleotide repeated adjacent to each other (http://en.wikipedia.org/wiki/Tandem_repeat). They are presumed to occur frequently in genomic sequences, for instance, they comprise at least

10% of human genome and are main cause of many genetic human diseases (Macleod *et al.*, 1999). They equally play a salubrious role in gene regulation by interacting with transcription binding factors, altering chromatin or acting as protein binding sites, and apparently participating also in the development of immune cells (Benson, 1999). The repeats can be of short or long pattern sequences. Micro-satellites are short tandem repeats ranging from 2-7 bp, while mini-satellites are composed of tandem repeats of 8–100 bp sequences (http://www.web-books.com/MoBio/Free/Ch3G1.htm). These are generally known as variable number of tandem repeats (VNTRs) if the repeat number is variable or unknown. Micro-satellite and mini-satellite sequences occur both within coding regions of the genome, including within antigenic sequences (Macleod *et al.*, 1999) and also in the IGRs and introns (Oura *et al.*, 2003). A previous study observed that depending on species, repeat regions can be perfectly conserved in sequence (Benson, 1999), but not in length, as the tract is subject to constant expansions and contractions, resulting in length variation even between chromosomes in the same cell. The extent of such expansions and contractions is much less in species with short tracts (Barry *et al.*, 2003). Extensive knowledge about tandem repeats in terms of copy number, pattern size, and mutational history among other aspects, however, has been limited by the inability to detect especially the long sequence patterns mainly due to difficulty in spotting them.

As a step towards both population and comparative studies in *Theileria* genus, a previous study reported a set of several polymorphic molecular markers (micro- and mini-satellites) distributed across the *T. parva* genome and located in both protein coding and non-coding regions (Oura *et al.*, 2003). In *Theileria*, these marker sequences are often scattered internally (interstitially located) within the genomes and often in tandem repeats (Bishop *et al.*, 1998). These repeat sequences have also been reported to have linkage disequilibrium and extensive genotypic diversity at their loci (Odongo *et al.*, 2006). While some satellite sequences, such as those in IGRs are likely to be selectively neutral (stochastic), others are probably under selective pressure, particularly those exposed to the immune system in which a nucleotide mutation may or may not result in a change in the amino acid sequence due to gene code redundancies and codon bias (Oura *et al.*, 2003). The authors were in support of different types of mechanisms of mutation operating within the micro-satellites and mini-

satellites. Micro-satellites, in general, show high rates of mutation as a result of replication slippage and point mutations with mismatch-repair deficiencies. Mini-satellites, on the other hand, generally show an even higher rate of variation, probably due to their mechanism of mutation that includes both meiotic crossing-over and replication slippage (Li *et al.*, 2004b). The mutation frequency in both classes is typically higher in IGRs and introns than that of base substitutions in ORFs encoding proteins (Li *et al.*, 2004b). Because these markers have the advantage of being typically widely dispersed and often selectively neutral, coupled with the high rates of mutation, they have high levels of polymorphism at the VNTR loci making them ideally suited for high resolution molecular fingerprinting of pathogen isolates (Bishop *et al.*, 2009). They have been used extensively in population genetic analyses, for example, in defining population structures, geographical sub-structuring and inbreeding levels in *P. falciparum*, in analyzing the *Babesia* genome (Brayton *et al.*, 2007), in characterizing *Babesia* isolates (Beck *et al.*, 2009), and in defining genotypic diversities of *Theileria parva* field isolates (Oura *et al.*, 2005; Odongo *et al.*, 2006).

The *T. parva* telomere-associated DNA has been reported to contain short tracts of repetitive sequences unlike the evidence in *P. falciparum* with long repeats. The potential CDS are often located at least 3kb of the telomeric repeats, and notably, over 70% of these genes are directly adjacent to at least three *T. parva* telomeres, as ORFs (Bishop *et al.*, 2000). These telomeric and sub-telomeric regions of protozoan genomes have been shown to encode multi-copy-polymorphic gene families which are known to be central to several aspects of parasite–host interaction, particularly pathogenesis and antigenic variation (Bishop *et al.*, 2000). Phenotypically, these result in variant surface proteins, protein modification, stabilization, degradation, targeting, sorting, translocation, and other protein fate-related functions (Carlton *et al.*, 2002). The structural importance of the location of these multi-copy-polymorphic gene families at the sub-telomeric regions is perhaps to allow rapid evolution of surface or secreted proteins (mainly by ectopic recombination and panmixia) which are crucial in the survival of these intra-cytoplasmic parasites (Bishop *et al.*, 2000). This observation is true for the *T. parva* despite the absence of long telomeric repeats in contrast to those observed in *P. falciparum* and *Saccharomyces cerevisiae* (Bishop *et al.*, 2009). The *T. parva* Marikebuni clone has been reported to lack a significant

set of sub-telomeric repeats altogether (Sohanpal *et al.*, 1995), substantially protracting the genome. However, it is not clear if these set of repeats play a role in virulence because, obscurely, the parasite induces lethal infections in naive mammalian hosts. The expression of the ORFs located adjacent to the telomeres has been reported to be modulated by telomere positional effect - also known as transcriptional silencing in *S. cerevisiae* and in trypanosomes (Bishop *et al.*, 2000). However, sub-telomeric gene expression in *T. parva* is apparently not completely repressed by telomere positional effects (Bishop *et al.*, 2000; Bishop *et al.*, 2009), possibly because of co-transcriptional regulation of tandem gene sets as already mentioned above.

The rapidly evolving gene family known as *T. parva* repeats *(Tpr)* is reported as widely dispersed within the genome with a complex domain especially at the 3' end of genes. These are isolate-specific and are used in genotyping of the isolates (Gardner *et al.*, 2005). These satellites are however, neither in tandem nor telomeric, but rather interstitially scattered in genome with highly conserved ORFs on chromosome 3 (Gardner *et al.*, 2005). The gene homologues of *Tpr* that are predominantly located at the sub-telomeres in *Plasmodium* include the *vir* (*P. vivax*), *var, rifins*, and *stevors* (*P. falciparum*) and *yir* (*P. y. yoelii*), whose products have been shown to be involved in antigenic variation, host cell invasion, and host immunologic evasion (Carlton *et al.*, 2002); and in *T. annulata*, the *Tar* family are dispersed within the genome in all the four chromosomes but lack the tandem arrays (Bishop *et al.*, 2009). The *p67* antigen which has variously been reported to be able to prevent the sporozoites from invading the host lymphocytes is also notoriously known to have hyper-variable tandem repeats with varied flanking regulatory sequences, especially upstream (Brayton *et al.*, 2007) and predominantly get transcribed in the schizont stage (Bishop *et al.*, 2009).

2.4 Codon usage

Codons are genetic codes that transfer information encoded in nucleic acids to proteins. The codons that code for the same amino acid are referred to as synonymous codons, and they are used at relatively different frequencies during translation (Grantham *et al.*, 1980; Sharp *et al.*, 1988). The genetic code is necessarily redundant with most amino acids

encoded by several synonymous codons (Sharp and Li, 1987). In many genomes, some codons are favored over others by selection likely because they are translated more efficiently and accurately, a phenomenon referred to as codon usage bias (Sharp *et al.*, 1988). Selection for the use of such favored codons is stronger for genes that are more highly expressed, such as ribosomal genes or translation elongation factors, exclusively exhibiting very high levels of codon bias (Sharp *et al.*, 1988). The identity of selectively favored codons varies among species including among strains and clones of a species (Sharp *et al.*, 1988; Hershberg and Petrov, 2009). For example, the favored codon for leucine in *Escherichia coli* and *Drosophila melanogaster* is CTG, in *Bacillus subtilis* TTA, in *Saccharomyces cerevisiae* TTG, and in *Saccharomyces pombe* CTT (Sharp *et al.*, 1988). Though an early research set some rules governing the identities of favored codons in different organisms (Sharp *et al.*, 1988) and codon usage in *Cryptosporidium parvum* (Grocock and Sharp, 2001), reports in *Theilerians* is lacking partly due to inavailability of annotated genomes and partly due to complicated analysis process. Codon bias has been a long recognized and long studied biological phenomenon, yet several basic questions regarding codon usage in ORFs containing VNTRs remain unresolved. Elucidation of the nature and variation of codon usage in any genome can yield insights into various aspects of molecular evolutionary processes in addition to basic aspects of the biology of the species, as well as providing a knowledge base for the interpretation of genome sequence data.

CHAPTER THREE
MATERIALS AND METHODS

3.0 Introduction

The current study was based on the nuclear genomes of *T. parva* Marikebuni and *T. parva* Muguga using *in silico* (computer based stand-alone prediction programs and web-based bioinformatics tools) approaches. The *T. parva* Marikebuni (piroplasm clone 3292, bled from animal BJ253, 27.11.1992) genome was recently sequenced by International Livestock Research Institute (ILRI) Bioinformatics group (un-published data) using whole-genome shotgun (WGS) titanium 454 technique (commonly known as Roche 454) and comparatively assembled using *T. parva* Muguga as template genome via AMOS Comparative Assembler (AMOS-Cmp) pipeline as described elsewhere (Pop, 2009). The technique involved shearing the whole genome into short, adapter-flanked fragments (reads) that are immobilized on a surface (28 µm beads) and simultaneously pyro-sequenced from both ends before being assembled using throughput-assemblers. The Roche 454 technique was particularly chosen because it produces longer reads (about 250-400 bp) that give better assemblies as compared to other new generation sequencing techniques like Illumina/Solexa and Helicos that produce short reads as low as 25-50 bp (Mardis, 2008; Shendure and Ji, 2008; Lister *et al.*, 2009; Pop, 2009). The *T. parva* Muguga genome was retrieved from the GenBank database with accession number AAGK00000000 (http://www.tigr.org/tbd/e2k1/tpa1) and used as template in the sequencing-assembling and annotation processes. The genomic analysis was based on micro-satellites (ms) and mini-satellites (MS) and their flanking primer pairs as previously reported (Oura *et al.*, 2003). Various bioinformatics tools were used as mentioned within the text but the Artemis Comparison Tool (ACT) (Carver *et al.*, 2008), release 9.0, was the principal software used to visualize, annotate and analyze the genomes.

3.1 Annotation and Curation of protein-coding DNA sequences (CDS)

The just assembled nuclear *T. parva* Marikebuni genome, as query file, was mega-blasted against the GenBank nuclear *T. parva* Muguga genome (Gardner *et al.*, 2005) as reference file, to assess their sequence similarities and identities, and confirmed via MUMmer-plots (Delcher *et al.*, 1999). The MUMmer program aligns whole genome sequences, often millions of base-pairs of closely related organisms by combining three algorithms (suffix trees, the longest increasing subsequence [LIS], and Smith-Waterman alignment) in finding maximum unique matching sequences (Delcher *et al.*, 1999). PASA (program to assemble spliced alignments) was used to predict all the protein-coding DNA sequences (CDS) in the *T. parva* Marikebuni genome from the *T. parva* Muguga protein sequences that were used as ESTs (Expression Sequence Tags) (Haas *et al.*, 2003; Pop and Salzberg, 2008). The output was written in GFF (Generic File Format) type format as a comparison tab file. For each of the four chromosomes of the genomes, the Marikebuni genbank (.gb) files were up-loaded onto ACT as a query file, followed by the comparison tab file, then a third file of *T. parva* Muguga annotated template (.gb) file to automatically align and visualize the CDS features. The features were manually checked for splice-point nucleotide demarcations for exon/intron boundaries and open reading frames (ORFs) shifts between the genomes (Haas *et al.*, 2003), before being committed as permanent records in the query file using the ACT saving option under file menu. The un-captured features un-captured/ or partially captured by the PASA soft-ware as maximal matches together with their feature identifiers were manually curated using ACT color matches as guide (**Figure 2**). The genomic CDS feature overviews for all the chromosomes in both genomes were retrieved from the ACT and analyzed. The CDS sequences were then retrieved and translated to their respective amino acid sequences. The stand-alone Interproscan software hosted at http://www.hpc.ilri.cgiar.org/tools was used to search Pfam, (a database of protein families that includes their annotations and multiple sequence alignments), Tigr, (a protein database developed and maintained by The Institute of Genomic Research, at the J. Craig Venter Institute), and superfamily (a database of structural and functional annotation of all proteins and genomes) databases generated using Hidden-Markov Models, (HMM)

algorithms] for gene structure protein-domains and families including known gene ontologies (GO).

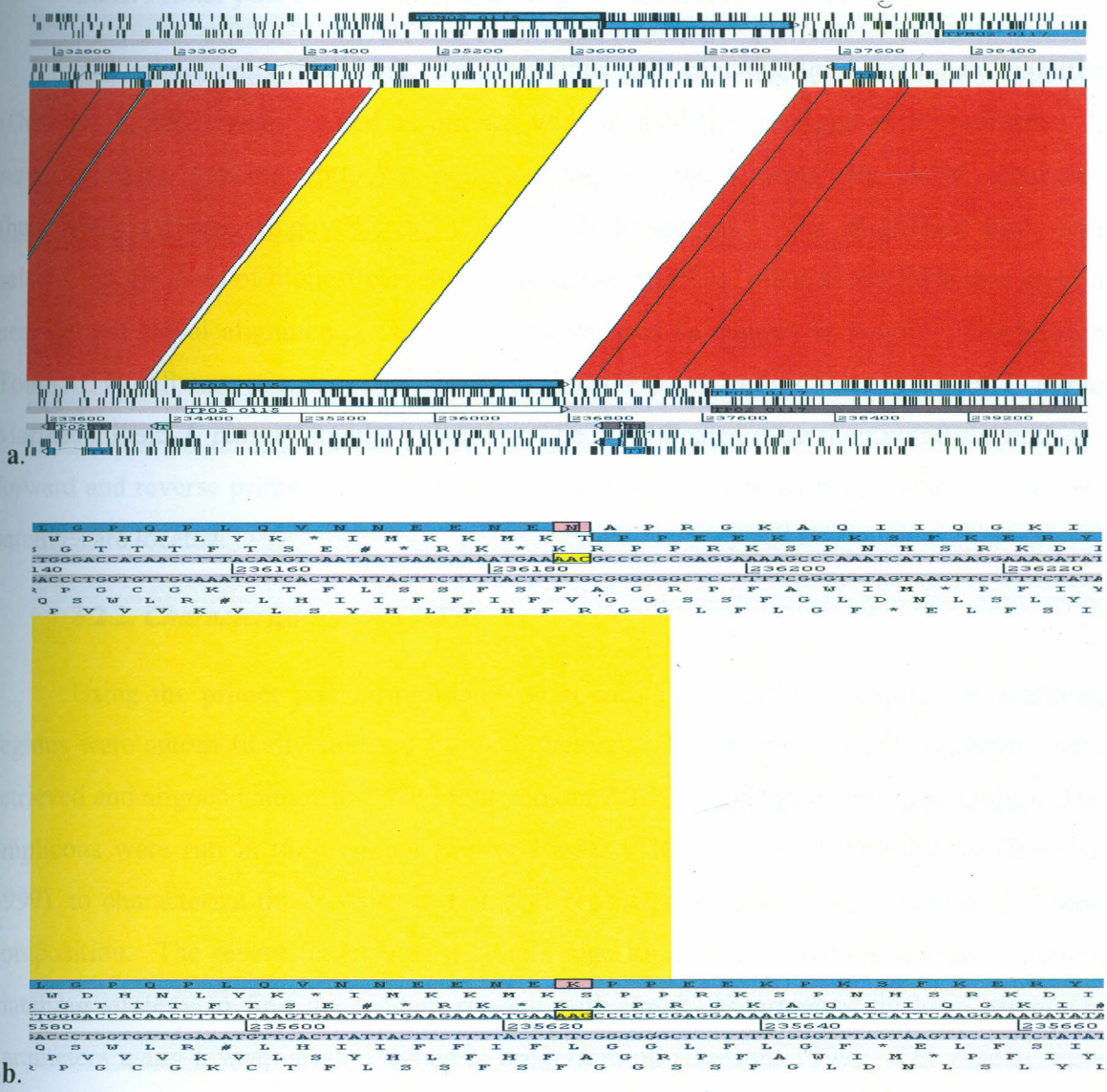

Figure 2: **Screenshot of a section of chromosome 2 ACT visualization for annotation process.** The horizontal panels above with arrow-ended blue boxes show *T. parva* Marikebuni genome and those below show *T. parva* Muguga genome. The blue boxes represent the CDSs annotated. In **(a),** the yellow colored part indicates a region of the genomes selected for annotation (for example locus 0115 showing yellow/white boundary of a spliced-junction in the Marikebuni genome and not in the Muguga genome). The red parts are other un-selected matching regions. In **(b),** the same locus 0115 zoomed in to show the nucleotides with the splice-point AA/GC indicated by pink color.

3.2. Variable Number Tandem Repeats (VNTRs)

3.2.1. *Primer pair co-ordinates and the genomic loci of VNTRs*

The primer pairs that flank micro-satellites and mini-satellites previously reported (Oura *et al.*, 2003) were blasted as queries without modification against the published *T. parva* Muguga genome (http://www.tigr.org/tbd/e2k1/tpa1) using BLASTN algorithm (http://www.ncbi.nlm.nih.gov/BLAST/) with default settings. The alignment hits were selected on the basis of highest percentage similarity matches, often above 80% that gives a near perfect global alignment. The primers were used as queries in Artemis Comparison Tool (ACT) software (Carver *et al.*, 2008) to find the primer co-ordinates of both *T. parva* Marikebuni and *T. parva* Muguga genomes. The co-ordinates were then noted for both the forward and reverse primers and used to calculate expected amplicon size within which the satellites are located.

3.2.2. *Characterization of VNTRs*

Using the primer pair co-ordinates as queries in the ACT software, the spanning regions were automatically selected for each genome. The primer-flanked amplicons were retrieved and aligned using ClustalW2 tool version 2.0.12 (http://www.ebi.ac.uk/tools/). The amplicons were run in the Tandem Repeat Finder (TRF) version 4.03 software (Benson, 1999), to characterize the VNTRs in terms of repeat pattern, size, copy number and base composition. The repeat finder uses a *k*-tuple algorithm coupled with a stochastic pattern matching strategy to find the tandem repeats with Smith-Waterman style local alignment parameters using a wrap-around dynamic programming (Benson, 1999). The repeats were then selected on the criteria of short repeat pattern with high copy number and high percentage match hit. Shorter repeat patterns are captured with high algorithmic accuaracy and precision, and the higher the copy numbers, the more polymorphic the repeat. When the adjacent repeat patterns are analyzed, a consensus pattern is derived from the smallest of them and used to define a report threshold. So the higher the percentage the stronger the repeat patterns match over all. These were counter-checked against the panel of satellites previously reported (Oura *et al.*, 2003). The TRF results were incorporated into the

ClustalW2 alignment files to indicate the exact positions of the VNTRs in the primer-flanked amplicons (**Appendix 1**).

3.2.3. *VNTR precise genomic loci*

The indices of the TRF repeats together with the amplicon co-ordinates were used to derive the precise genomic loci of the VNTRs in both genomes. The exact genomic location of the VNTR patterns were noted for three categories: coding open reading frames (ORFs), introns, or intergenic regions (IGRs). The bases and amino acids of VNTR-containing ORFs were selected and retrieved from the ACT software. These were aligned using the ClustalW2 tool and used for further analysis. The peptide domains and motifs of VNTR-containing ORFs were retrieved from the fore-mentioned database outputs and used to infer the protein motif hits, relative to the VNTR locations.

3.3.1 Codon Usage analysis of coding ORFs bearing the VNTRs

3.3.1. PCR amplifications and sequencing of gapped ORFs containing VNTRs

Due to assembly insufficiency, the coding ORFs containing VNTR patterns with un-sequenced gaps were identified and retrieved for manual PCR sequencing. In this, primers flanking the gapped regions were designed using primer3 software hosted at the ILRI high performing computer (http://hpc.ilri.cgiar.org/tools/emboss/) (**Appendix 2**). The primers were manufactured by Bioneer Inc. (http://us.bioneer.com/). Primer DNA were reconstituted as per the manufacturer's instructions to make 100 pmol/µl, and were run on 2% agarose gel with 2µl ethidium bromide stain, electrophoresed in 0.5x TAE electrolyte, visualized and photographed under ultraviolet light to confirm them. Bromothymol blue 1X loading die was used at 4 µl/µl of PCR reaction mix. The PCR amplification conditions and results were as indicated in **Appendix 3 (a)**. The PCR products were purified using QIAquick PCR Purification Kit (250) protocol as per manufacturer's instructions (http://francois.schweisguth.free.fr/protocols/QIAquick_PCR_Purification_Kit). The PCR products were precipitated using 30% PEG/$MgCl_2$ at 0.5 volume of each sample, vortexed to mix and centrifuged at 15rpm at 20°C for at least 30 minutes then quantified by gel electrophoresis strength (**Appendix 3 b.)**. The supernatant was recovered in 1.5ml ependorf

tubes and stored at -20°C. The DNA pellets were then dissolved in 10µl de-ionized reagent water. Of each DNA product, 1µl was used for electrophoresis analysis and about 0.1µl used in nanodrop spectrophotometry absorbance (**Appendix 4**). The PCR products containing at least 50ng/µl DNA with their respective primer concentrations at 100ng/µl were sent for sequencing using the 454 Sequencer (Genome Sequencer FLX Titanium Series, © 1996-2010 Roche Diagnostics Corporation, Roche, Switzerland). The sequencing results were assembled using Staden DNA assembler – an open source software version 2.0 (http://staden.sourceforge.net/overview.html) and manually edited before being inserted into their respective ORF loci.

3.3.2 Codon usage analysis

Codon usage analysis was implemented via a stand-alone CodonOptTable, Bio::Tools::CodonOptTable software version 0.07 (http://search.cpan.org/~shardiwal/Bio-Tools-CodonOptTable-0.07/lib/Bio/Tools/CodonOptTable.pm) and statistical analysis done via command line CodonO webserver (http://www.sysbio.muohio.edu/CodonO/). The CodonOptTable program does codon bias analysis by measuring two parameters, Relative Synonymous Codon Uses (RSCU) and Relative Adaptiveness of a Codon (RAC) based on previous studies (Sharp and Li, 1987). The RSCU value for a codon is simply the observed frequency of that codon divided by the frequency expected under the assumption of equal usage of synonymous codons for an amino acid (Ikemura, 1985). In the absence of any codon usage bias, the RSCU value would be 1.00. The RAC is calculated based on RSCU value, the frequency of that codon as compared to the frequency of the optimal codon for that amino acid. Intrinsically, the conversion of codon usage values to RSCU and RAC values helps normalize the otherwise non-normal data across genes and genomes. It makes the codon usage value independent of amino acid composition of the sequences and identifies when a codon is being used more frequently than expected and when it is being used less frequently than expected.

CHAPTER FOUR

RESULTS

4.1 Annotation of *T. parva* Marikebuni genome

The nuclear genome of *T. parva* Marikebuni was partially sequenced and assembled into four chromosomes ranging between 1,485,980 nucleotide pairs in chromosome 1 and 2,536,855 nucleotide pairs in chromosome 2. Chromosome 3 was assembled into two contiguous sequences due to inability of the assembly machinery to resolve long stretches of mono-nucleotide repeats. The sequencing and assembling of the genome gave over 90 % coverage with significant gaps falling even within intragenic regions. The resultant annotated genome showed a perfect synteny with *T. parva* Muguga, especially in the coding regions. A total of 3937 and 4005 nuclear CDS were present across the four chromosomes of *T. parva* Marikebuni and *T. parva* Muguga genomes, respectively. The *T. parva* Marikebuni CDSs were assigned unique feature identities as TPM0n_xxxx (where n denotes chromosome number while x denote locus number) to distinguish them from the same nomenclature used for *T. parva* Muguga (i.e. TP0n_xxxx) (Gardner *et al.*, 2005). The genomes have subtle variation at nucleotide level but have high similarity at amino acid level mainly due to synonymous codon redundancies. These are shown by the megablast alignments (**Figure 3 i-iv**) and best-of fit MUMmer-plots (**Figure 4**). The MUMmer-plots present unique matches at nucleotide level via Nucmer program. The variations at base level are mainly due to insertion-deletions and substitutions between the genomes. As indicated by the blue diagonal lines that cross each other in Figure 3, some nucleotide sequences remain conserved but inverted.

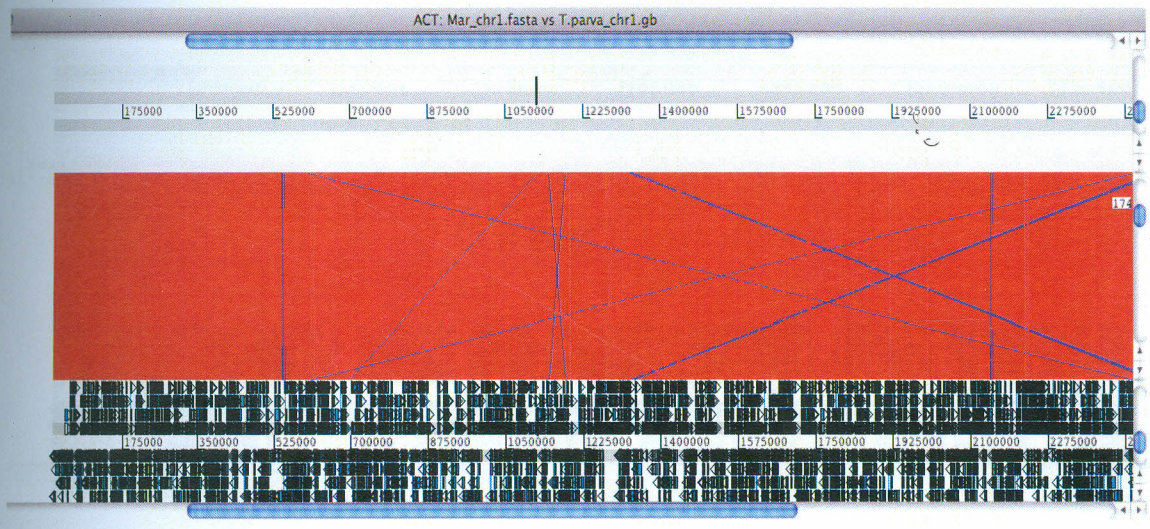

Figure 3 i. Chromosome 1. Megablast alignment snapshots of *T. parva* Marikebuni against *T. parva* Muguga from ACT software. The diagonal lines crossing each other indicate nucleotide sequence inversions. The red panel show matching regions along the diagonals. The *T. parva* Marikebuni genome is placed on top of the red coloured panel as a blank linear DNA sequences while the *T. parva* Muguga genome is laid below the red coloured panel with condensed annotations.

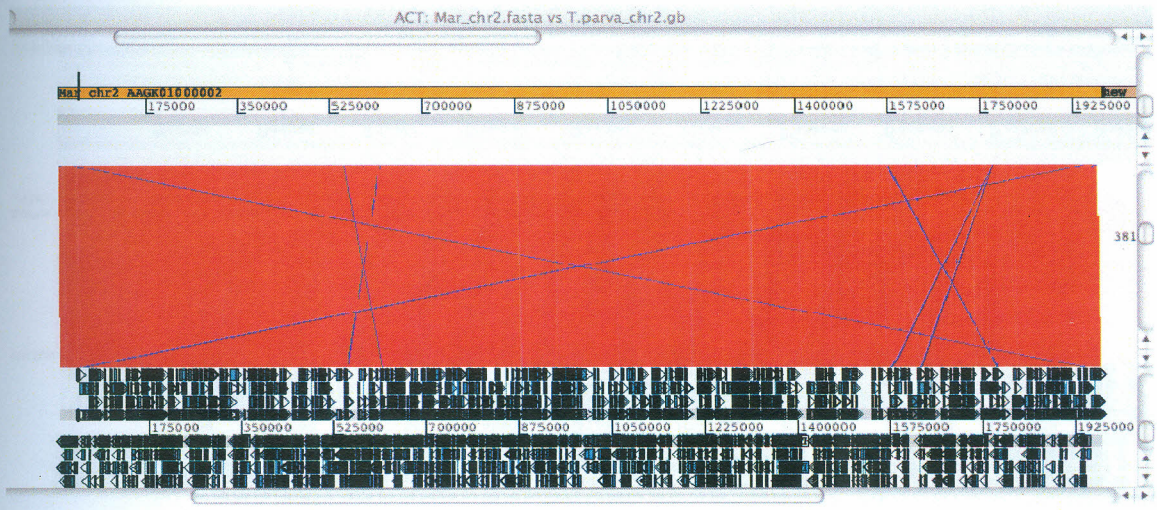

Figure 3 ii. Chromosome 2. Megablast alignment snapshots of *T. parva* Marikebuni against *T. parva* Muguga from ACT software. The diagonal lines crossing each other indicate nucleotide sequence inversions. The red panel show matching regions along the diagonals. The *T. parva* Marikebuni genome is placed on top of the red coloured panel as a blank linear DNA sequences while the *T. parva* Muguga genome is laid below the red coloured panel with condensed annotations.

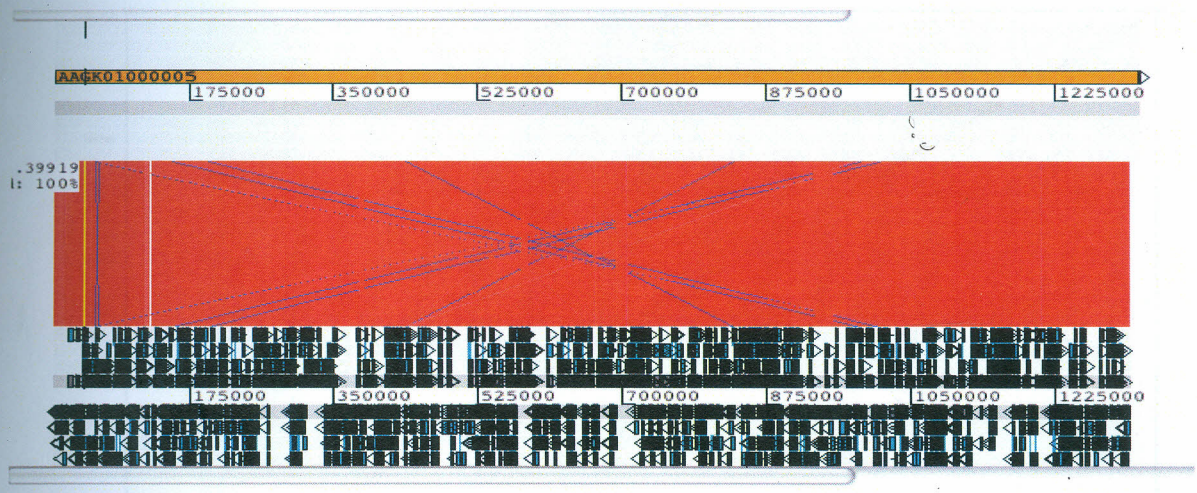

Figure 3 iii. Chromosome 3/530. Megablast alignment snapshots of *T. parva* Marikebuni against *T. parva* Muguga from ACT software. The diagonal lines crossing each other indicate nucleotide sequence inversions. The red panel show matching regions along the diagonals. Only contig 530 of chromosome 3 is shown. The *T. parva* Marikebuni genome is placed on top of the red coloured panel as a blank linear DNA sequences while the *T. parva* Muguga genome is laid below the red coloured panel with condensed annotations.

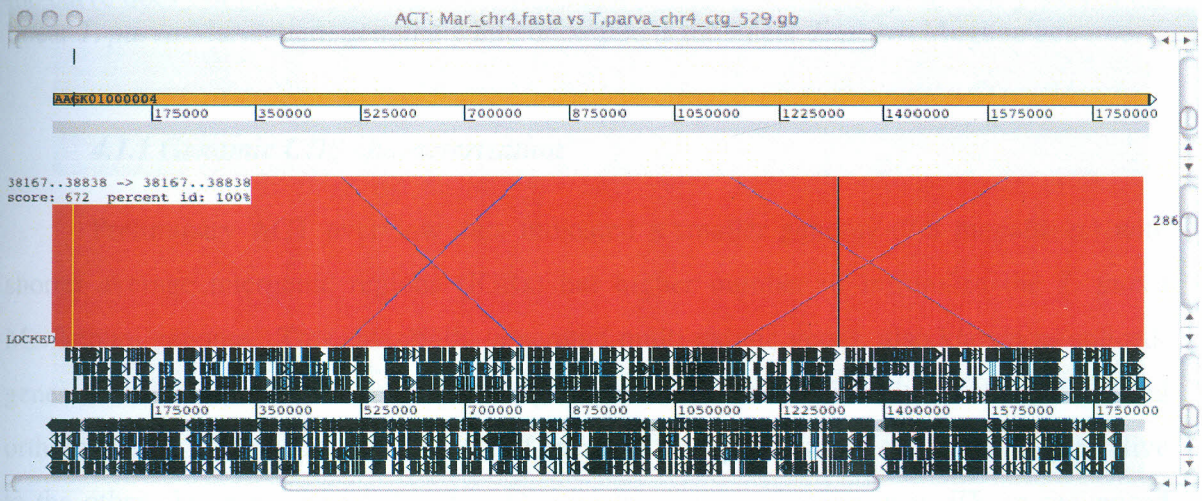

Figure 3 iv. Chromosome 4. Megablast alignment snapshots of *T. parva* Marikebuni against *T. parva* Muguga from ACT software. The diagonal lines crossing each other indicate nucleotide sequence inversions. The red panel show matching regions along the diagonals. The *T. parva* Marikebuni genome is placed on top of the red coloured panel as a blank linear DNA sequences while the *T. parva* Muguga genome is laid below the red coloured panel with condensed annotations.

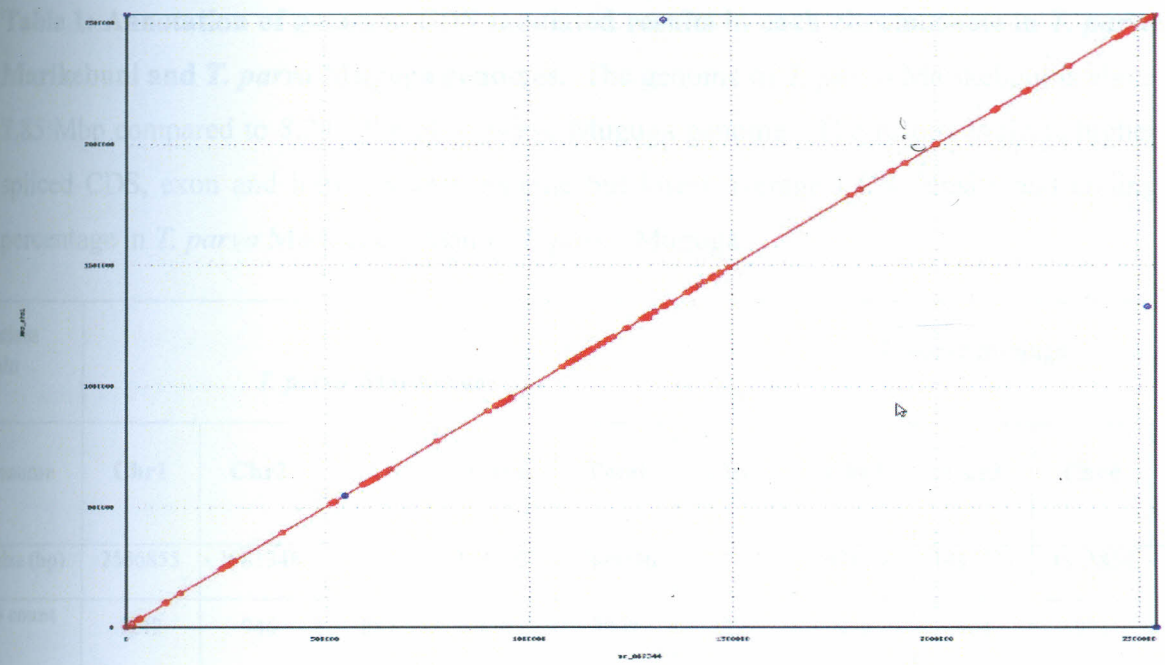

Figure 4: **Best-of-fit Nucmer MUMmerplot alignment of chromosome 1 nucleotides of *T. parva* Marikebuni and *T. parva* Muguga.** The nucmer plot shows a good synteny between the two genomes though not uniformly because of sequencing gaps. This is why the curve does not have equal thickness but shows a continous linear plot. On y-axis is the linear *T. parva* Marikebuni genome while on x-axis is the linear *T. parva* Muguga genome.

4.1.1 Genomic CDS characterization

Preliminary ACT overview data indicated a smaller genome (7.8 megabases) with shorter average intergenic regions (IGRs) but higher number of introns in the *T. parva* Marikebuni genome. This essentially generated multi-exonic genes, also known as spliced-genes. It is unclear if these multi-exonic genes retain their functionality as the less-spliced orthologues in *T. parva* Muguga. Most of these spliced-genes are in chromosome 1 relative to the other chromosomes (**Table 1**). Despite the genome having over 4% un-sequenced gaps within the coding regions, the results indicated a comparable gene density at 0.473 against 0.487 per kilo base pair and a coding percentage of 65.9 versus 68.4 for *T. parva* Marikebuni and *T. parva* Muguga, respectively.

Table 1: Annotation of genomic CDS tabulated results in each chromosome in *T. parva* Marikebuni and *T. parva* Muguga genomes. The genome of *T. parva* Marikebuni is about 7.85 Mbp compared to 8.24 Mbp of *T. parva* Muguga genome. The results indicate higher spliced CDS, exon and intron counts genome but lower average CDS density and coding percentage in *T. parva* Marikebuni than in *T. parva* Muguga.

Organism strain		*T. parva* Marikebuni					*T. parva* Muguga				
Chromosome		Chr1	Chr2	Chr3	Chr4	Total	Chr1	Chr2	Chr3	Chr4	Total
Genome size (bp)		2536855	1981348	1485980	1845107	**7849290**	2540030	1971884	1887728	1835834	**8235476**
Total CDS count (bp)		1212	946	882	897	**3937**	1222	958	904	921	**4005**
CDS count (bp)	Spliced	1039	766	710	750	**3265**	951	687	645	694	**2977**
	Non-spliced	173	180	172	147	**672**	271	271	259	227	**1028**
Exons count		5162	3468	3457	3745	**15832**	4616	3165	3137	3498	**14416**
Av. Exon length (bp)		332.3	382.4	357	332.9	**352.32**	377.8	425.9	407.1	366	**394.2**
Introns count		3950	2522	2575	2848	**11895**	3394	2207	2233	2577	**10411**
Av. Intron length (bp)		77.8	86.0	98.45	75.0	**73.14**	90.3	96.4	109.6	80.9	**94.3**
Average CDS density (genes/kbp)		0.477	0.477	0.463	0.486	**0.473**	0.481	0.485	0.482	0.501	**0.487**
Average Coding (%)		67.6	66.9	63.75	67.5	**65.9**	68.6	68.3	67.1	69.7	**68.425**

Computation for intergenic regions (IGR) length (bp): as total genome (TG) size minus sum of exons and introns in bp plus 1.

IGR = TG – ((Exon count * av. Exon length) + (Intron count * av. Intron length)) (bp)
i.e.
T. parva Marikebuni:
Average IGR length: total IGR length divided by total CDS count plus one
= (1401359.46 / 3937) + 1

= 357bp

T. parva Muguga:
Av. IGR length:
= (1570931.5 / 4005) + 1
= 392.24 + 1
=393 bp

The nucleotide composition was compared between the two genomes, with results showing lower values across the four chromosomes for *T. parva* Marikebuni. The genomic GC content of *T. parva* Marikebuni was equally lower at 32.5% compared to 34.06% in *T. parva* Muguga (**Table 2**). However, the variations observed here could narrow if the Marikebuni genome was cleaned of the gaps.

Table 2: Overall nucleotide compositions of genomic CDS in both genomes expressed as percentage values. *Theileria parva* Marikebuni has 4.12% un-sequenced regions represented here by N. Both genomes are AT-rich. Generally, the *T. parva* Marikebuni genome has lower nucleotide content compared to the *T. parva* Muguga genome, suggestive of a contracted genome.

Bases (%)	*T. parva* Marikebuni						*T. parva* Muguga					
	Chr 1	Chr 2	Chr 3/350	Chr 3/531	Chr 4	Average	Chr 1	Chr 2	Chr 3/530	Chr 3/531	Chr 4	Average
Adenosine (A)	32.03	31.6	31.34	30.4	31.77	**31.568**	32.96	32.74	32.61	33.62	32.9	**32.928**
Cytosine (C)	16.63	16.67	16.37	15.12	16.42	**16.366**	17.12	17.2	17.05	16.99	17.03	**17.093**
Guanosine (G)	16.54	16.62	16.15	15.02	16.32	**16.266**	17.03	17.18	16.82	16.63	16.92	**16.964**
Thymine (T)	32.02	31.71	32.16	29.77	31.93	**31.281**	32.87	32.87	33.51	32.75	33.14	**33.003**
N	2.76	3.38	3.95	9.67	3.53	**4.12**	0	0	0	0	0	**0**
GC %	33.18	33.29	32.53	30.15	32.75	**32.64**	34.16	34.38	33.87	33.63	33.95	**34.06**

4.1.2 CDS Protein motifs and domains

The comparative annotation predicted a total of 3937 CDS across the four chromosomes of *T. parva* Marikebuni (**Table 1 and Table 3**) excluding whole homologues that fell within un-sequenced stretches. The study, through interproscan database searches with intrinsic cut-offs, was able to assign a total of 2231 putative domains and families to the CDSs of *T. parva* Marikebuni. Of this number, 1407 (representing 63.07% of total) had gene ontology (GO) hits; 824 (36.93%) reported domains with unknown functions while 1641 (41.68%) CDS had no domain hit-matches from the three databases searched. In comparison to the *T. parva* Muguga genome, a total of 2312 CDS (57.73%) had motif homologs of which 2103 (52.51%) had GO identities while 209 (5.22%) had protein motifs of unknown functions (**Table 3**).

Table 3: InterProscan whole genome protein domain prediction statistics of *T. parva* Marikebuni compared to *T. parva* Muguga. The *T. parva* Marikebuni indicates more CDS with unknown functions compared to the Muguga genome. The asterix (*) indicate the combined CDS statistics of the two contiguous sequences 530 and 531 of chromomsome 3 but excludes an AT-rich repeat region that should be containing *Tpr* (*T. parva* repeat) hyper-polymorphic gene family.

Organism	***T. parva* Marikebuni**					***T. parva* Muguga**				
Chromosome	**TPM01**	**TPM02**	**TPM03***	**TPM04**	**Total**	**TP01**	**TP02**	**TP03***	**TP04**	**Total**
Total CDS number	1212	946	881	898	**3937**	1222	958	904	921	**4005**
CDS with protein/domain (hits)	705	546	466	514	**2231**	742	564	475	531	**2312**
CDS with gene ontology (GO) hit	442	332	305	328	**1407**	688	505	426	484	**2103**
CDS with Unknown protein/domain functions	263	214	161	186	**824**	54	59	49	47	**209**
CDS with No protein/domain hits	478	389	405	369	**1641**	480	394	429	390	**1693**
CDS Un-accounted for	29	11	10	15	**65**		-	-	-	-

The other regions of interest included the species-specific telomeric hyper-variable gene families that map adjacent to the telomeres, the adenosine triphosphate binding cassettes (ABC) transporter genes that demarcate the housekeeping from hyper-variable species-specific genes and the antigenic sporozoite surface protein, p67, reviewed in a previous study (Bishop *et al.*, 2009). A good synteny was observed in both sub-telomeric regions that encode highly divergent paralogous gene families as previously reported for *T. annulata* (Pain *et al.*, 2005) and centromeric regions that encode housekeeping genes (Shiels *et al.*, 2006). For example, the ABC-transporter genes indicated high level of conservation in terns of structural location and composition in both these *Theileria parva* genomes. However, full report of these regions is beyond the scope of this study.

4.2 Characterization of VNTRs

In the *T. parva* Marikebuni genome, as depicted by **Table 4,** the flanking primer pairs searched in the ACT software did not find matching hits for mini-satellites MS4_forward, MS8_forward and MS9_forward in chromosome 1; MS12_reverse, MS14_reverse, MS16_forward and both pairs of MS8 in chromosome 2; micro-satellites ms9_reverse and ms10_reverse, both flanking pairs for mini-satellites MS23 and MS27 in chromosome 3; and mini-satellites MS39_reverse, MS42_forward, MS43_forward, MS221/1_reverse and MS221/2_forward in chromosome 4. While searching for the primer pair co-ordinates in ACT, MS221/1_reverse was not found in the Muguga genome. The primer-flanked amplicons when run in the TRF version 4.03 software gave similar tandem repeats as those previously reported (Oura *et al.*, 2003) with subtle variations at repeat copy numbers and point mutations in repeat patterns. Micro-satellites ms3 and ms5, respectively in chromosome 1 gave variant repeat patterns TAGTAGAAG with 11.1 copy number and TTATTTTATAA with 5.8 copy numbers in both genomes as compared to tri-nucleotide patterns TAG with 35 copies and ATT with 26 copies in previous studies (Oura *et al.*, 2003). The repeat patterns reported here fit mini-satellite definition and seem to be as a result of duplication event which the previous version of TRF could not detect because of algorithm limitation. Mini-satellite MS2 had two patterns in the Marikebuni genome versus three

different patterns in the Muguga genome, all with differing copy numbers and variants to the pattern previously reported (Oura *et al.*, 2003). MS3 had an additional new confounding repeat pattern to those previously reported (Oura *et al.*, 2003) in both genomes, but the Marikebuni genome had exactly two copies of each lower than those in Muguga genome. MS4 reported a contracted 8-bp pattern with 20.1 tandem repeat in Muguga that seem truncated from previous PCR allelic pattern (Oura *et al.*, 2003). In chromosome 2, amplicons of ms7 and MS15 had no repeat patterns from the primer-flanked amplicons in Marikebuni but reported different patterns of AAT and AATTTAACAT with conserved copy numbers 30.3 and 11.7, respectively, in Muguga. Marikebuni genome had confounding new repeat patterns with lower copy numbers compared to Muguga in MS17 and MS19. Chromosome 3 amplicons of ms8_530 and MS31 gave no tandem repeat pattern in Marikebuni genome but the results in Muguga genome were consistent with previous reports (Oura *et al.*, 2003). Micro-satellite ms9_530 of Muguga genome returned a 5-bp pattern, TATAC, having 27.8 copies occurring together in same loci range with two other significant but confounding mini-satellite patterns, ACTATTAT and TATACCTATAC with 17.6 and 10.3 copy numbers, respectively. Similarly, mini-satellite MS23 gave a confounding 5-bp pattern ACTAT with 8.6 copies appearing in the same amplicon region with actual 10-bp pattern, GATAACAGTG, at 7.4 copies. MS22_530 reported two significant patterns of 11 bp each in both genomes while MS29_531 had two 33 repeat patterns in Marikebuni and one 32 repeat pattern in Muguga contrasting previous allelic PCR product of previous report of TAAGAGTAAAA at 20 copies (Oura *et al.*, 2003). In chromosome 4, mini-satellites MS32 and MS41 reported new repeat patterns, one extra each, in both genomes while MS38 had an identical new pattern, TATATACAATT, with 9.4 and 8.4 copy numbers in Marikebuni and Muguga, respectively. Mini-satellite MS46 in the Marikebuni genome had a very expansive amplicon with over 6 kilobase pairs (kbp) overlapping four distinct protein-coding DNA sequences (CDS). This satellite gave five tandem repeat patterns, 9-bp each, at different loci. Two of these located at amplicon index 153-230 and 6283-6333, were identical (i.e. TCAACCATA) but differed in copy number, 8.7 and 5.7, respectively. The Muguga genome had only one 9-bp pattern ortholog identical to the two above but higher copy number at 9.7. All the VNTR primer-flanked amplicon

sequences from the two genomes were retrieved and aligned against each other using ClustalW (version 2.0.12). The multiple sequence alignment of the VNTRs were combined with tandem repeat finder (TRF version 4.03) results to show regions of tandem repeats, repeat sequence patterns and their repeat copy numbers (see **Appendix 1).**

combination of BLASTN algorithm, ACT and TRF version 4.03 software results.

(a) Chromosome 1. Only reverse primer co-ordinates of mini-satellites MS4, MS8 and MS9 were located in *T. parva* Marikebuni genome from the ACT. Unexpectedly, micro-satellite ms3 reported a tandem repeat pattern size of 9 nucleotides in both genomes, which qualifies the definition of a mini-satellite. MS3 and MS817 were located in the same ORF, 0698.

Satellite	*T. parva* Marikebuni							*T. parva* Muguga						
	Amplicon coodinates	TRF VNTR index	VNTR pattern	VNTR size	Copy no.	VNTR Precise coodinates	Physical loci	Amplicon coodinates	TRF VNTR index	VNTR pattern	VNTR size	Copy no.	VNTR Precise coodinates	Physical loci
ms1	1181414..1181768	104-163	AAT	3	21	1181517-1181576	IGR 3' _0566	1183805..1184163	104-167	AAT	3	22	1183908-1183971	IGR
ms2	963836..964049	78-151	TATTATA	7	9.9	963913-963986	Intron _0469	966125..966309	75-130	TAT	3	19.7	966199-966254	Intron
ms3	1101252..1101598	172-280	TAGTAGAAG	9	11.1	1101423-1101530	Exon out frme_0530	1103398..1103708	136-244	TAGTAGAAG	9	11.1	1103533-1103642	Exon
ms4	1110607..1111034	75-371	AGTAT	5	57.8	1110682-1110977	Intron _0535	1112588..1113074	76-430	GTATA	5	68.6	1112663-1113017	Intron
ms5	1420021..1420186	58-121	TTATTTTATAA	11	5.8	1420078-1420141	IGR 5' _0679	1422131..1422297	59-122	TTATTTTATAA	11	5.8	1422189-1422252	IGR
MS1	1101742..1101972	96-176	ATAAATAAAAT	11	7.4	1101837-1101917	IGR 5' _0531	1103853..1104083	76-176	ATAAATAAAAT	11	7.4	1103928-1104028	IGR
MS2	1413168..1413519	108-211	TACACATT	8	11	1413275-1413378	Intron _0676	1415253..1415583	231-261	ATAATTGAGT	10	3.1	1415483-1415513	Intron
		117-201	ACACTATTTAC	11	8.3	1413284-1413368	Intron _0676		101-165	ATTTACACACTA	12	5.8	1415353-1415417	Intron
		x	x	x	x	x	x		116-180	TACACATTTTTC	12	5.8	1415368-1415432	Intron
MS3	1474567..1474316	58-189	TTATATACCAAA	12	12.7	1474510-1474379	Intron _0698	1477498..1477241	46-197	TTATATACCAAA	12	14.7	1477453-1477302	Intron
		73-186	TATACCAAAT	10	11.4	1474495-1474382	Intron _0698		61-194	TATACCAAAT	10	13.4	1477438-1477305	Intron
MS4	..1099574	incomplete	-	-	-	-	-	1102145..1101684	175-359	TAATACTA	8	20.1	1101871-1101787	IGR
MS5	1096946..1096666	60-191	ATATTAT	7	18.9	1096885-1096756	IGR 3' _0529	1099012..1098735	57-188	ATATTAT	7	18.9	1098956-1098825	IGR
MS6	939980..939850	61-108	ATTAAAATATA	11	3.8	939920-939873	IGR 5' _0460	941778..941392	49-256	AAATAATCATC	11	18.9	941730-941523	IGR
MS7	646808..646660	65-112	AACTATGTAAACAG TAACTGAA	22	2.2	646744-646697	_0316 Exon-frameshifts	648009..647636	57-306	GTAACTATAACTATGTA AACA	21	11.9	647953-647704	Intron/exon_0316
		65-119	AACTAGGTAAACA GTAACTGAA	22	2.5	646744-646690	_0316 Exon-frameshifts		x	x	x	x	x	x
MS8	..266613	incomplete	-	-	-	-	-	266891..266578	45-241	TTACACAGTA	10	19.7	266847-266651	Intron
MS9	..255829	incomplete	-	-	-	-	-	256399..256203	67-170	TATTATTAAC	10	10.4	256333-256230	Intron
MS10	243091..242845	35-183	TAAATAGTATA	11	13.5	243057-242909	Intron-s0120	243686..243429	35-194	TAAATAGTATA	11	14.5	243652-243493	Intron
MS817	1472627..1472476	38-112	ATTTTACACT	10	7.7	1472590-1472516	_0698 Exon	1475524..1475342	39-132	ATTTTACACT	10	9.7	1475486-1475382	Exon_0698
		x	x	x	x	x	x		39-146	ATTTTACACA	10	11	1475486-1475379	Exon_0698

b. Chromosome 2: Micro-satellite ms7 and mini-satellite MS15 in *T. parva* Marikebuni reported no tandem repeat pattern from the TRF software; mini-satellites MS12, MS14 and MS16 located only the reverse primer blast result in ACT while MS18 did not have primer co-ordinates located.

Satellite	*T. parva* Marikebuni							*T. parva* Muguga						
	Amplicon coodinates	TRF VNTR index	VNTR pattern	VNTR size	Copy no.	VNTR Precise coodinates	Physical loci	Amplicon coodinates	TRF VNTR index	VNTR pattern	VNTR size	Copy no.	VNTR Precise coodinates	Physical loci
ms6	1153466..1153646	28-133	AT	2	56.5	1153493-1153598	Intron_0570	1149076..1149256	28-133	AT	2	56.5	1149103-1149208	Intron
		34-122	TATATA	6	12.1	1153499-1153587	Intron		34-122	TATATA	6	12.1	1149109-1149197	Intron
ms7	1159872..1160046	nil					IGR	1155482..1155655	53--134	AAT	3	30.3	1155534-1155615	IGR 0573/0574
MS11	1230311..1230630	104-136	ACACACATTTAC	12	2.8	1230495-1230527	IGR_0602/0603	1224954..1225267	74-254	CACACTCTTTA	11	17.4	1225027-1225207	IGR
MS12	855299..855610	71-95	AAAATAAGGT	10	2.5	-	Intron_0423	852734..853045	122--237	TAAAATAAGCCTAGAA TAGAGG	22	5.3	852855-852970	Intron-compl
		122--174	TAAAATAAGCCTAG AATAGAGG	22	2.4									
MS13	828137..828461	206-293	TGTGTAAAA	9	9.4	828342-828429	in Exon-0413	826373..826653	162-249	TGTGTAAAA	9	9.4	826534-826621	Exon _0413
		x	x	x	x	x	x		69-129	TGTGTAAAAA	10	6	826441-826501	Intron_0413
MS14	1253824..1254520	113-173	TTTTACACAT	10	6.1	1253936-1253995	Exon_0615	1248020..1248716	113--577	TTTTACACAT	10	46.9	1248132-1248596	Exon
MS15	1798018..1797812	nil	-	-	-	-	-	1789213..1789008	54--165	AATTTAACAT	10	11.7	1789160-1789049	Intron_0881
		-	-	-	-	-	-		112--183	ATAATTTAT	9	7.8	1789102-1789031	Intron
MS16	1264298..1263926	73-144	TAACAAATATTAGT AAA	17	4.2	-	IGR _0619/0620	1258364..1257992	68--301	ACTAATATTTGTTATTT	17	13.8	1258297-1258064	IGR
		247-300	AAATATTGGTGAAT AAT	17	3.2									
MS17	798523..798162	46-146	TAACTGTGTAAA	12	8.4	798478-798378	Intron-0401	797000..796639	46-336	TAACTGTGTAAA	12	24.3	796955-796665	intron sense
MS18	-	absent	-	-	-	-	-	366175..365800	69--336	TAAATATGTG	10	28.2	366107-365840	intron-0178
MS19	355317..354911	149-340	TAATTAAAC	9	21.7	355169-354978	IGR 3'_0174	354329..354023	49-240	TAATTAAAC	9	22.7	354281-354090	IGR
		257-340	AATTAAAC	8	10.6	355061-354978	IGR		x	x	x	x	x	x
MS20	351941..351735	53--176	TGAGTTAGTAAC	12	10.3	351889-351766	Intron_0172	351074..350868	53-176	TGAGTTAGTAAC	12	10.3	351022-350899	Intron

only the forward primers in the *T. parva* Marikebuni genome. Micro-satellite ms8_530 and mini-satellite MS31_531 reported no tandem repeat pattern from the TRF software.

Satellite	*T. parva* Marikebuni							*T. parva* Muguga						
	Amplicon coodinates	**TRF VNTR index**	**VNTR pattern**	**VNTR size**	**Copy no.**	**VNTR Precise coodinates**	**Physical loci**	**Amplicon coodinates**	**TRF VNTR index**	**VNTR pattern**	**VNTR size**	**Copy no.**	**VNTR Precise coodinates**	**Physical loci**
ms8_530	202839..203111	nil	-	-	-	-	-	198667..198921	163-218	AT	2	28	198829-198884	IGR -0104 0105
ms9_530	1135199..1135460	26--204	TACCTATACTA(MS)	11	15.7	-	-	1124644..1124872	31-173	TATAC	5	27.8	1124674-1124816	Intron_0530
		143—200	TATACCATTA(MS)	10	5.6	-	-		19-172	ACTATTAT (MS)	8	17.6	1124662-1124815	Intron
		-	-	-	-	-	-		39-164	TATACCTATAC (MS)	11	10.3	1124682-1124807	Intron
ms10_531	17758..17475	nil	-	-	-	-	IGR 3'-0623/0624	17701..17418	121-172	AT	2	28	17581-17530	IGR -0623 0624
MS21_530	38675..38353	32-168	ATACTATT	8	17	38654-38508	Intron_0015	37058..36736	32-266	ATACTATT	8	29.3	37027-36793	Intron
MS22_530	192690..193289	46-172	TATTTTAGAGG	11	11.6	192735-192861	IGR 3'_0098	189267..189770	46-150	TATTTTAGAGG	11	9.6	189312-189416	IGR
		166-299	TATTTTAAGCA	11	12.5	192855-192987	IGR		144-203	TATTTTAAGCA	11	5.5	189410-189469	IGR
MS23_530	-	absent	-	-	-	-	-	640350..640514	20-62	ACTAT (ms)	5	8.6	640379-640411	Intron-rev_0309
	-	-	-	-	-	-	-		66-139	GATAACAGTG	10	7.4	640415-640488	Intron
MS24_530	675830..676137	29-68	TTTTCCAA	8	4.8	675858-675897	Intron-0321	668774..669071	29-117	TTTTCCAA	8	10.1	668802-668890	Intron-compl
		x	x	x	x	x	x		29-130	TTTTCCAAA	9	11.4	668802-668903	Intron-compl
MS25_530	999311..999619	140-277	TTATATAGTTAAGT	14	9.9	999450-999587	Intron_0476	990577..990884	27-276	TTATATAGTTAAGT	14	17.9	990603-990852	Intron
MS26_530	1159113..1159437	64-176	ATTCAATAA	9	13.1	1159176-1159288	IGR-5'_0540	1148073..1148396	63-178	ATTCAATAA	9	13.1	1148135-1148250	IGR
MS27_531	-	absent	-	-	-	-	-	454506..454714	40-148	TAATCAAATTAT	12	9.1	454545-454653	Intron-0840
MS28_531	73431..73041	110-151	TATTATTAACTACT	14	3	73322-73281	Intron-0654	73001..72611	39-244	TATTATTAACTA	12	16	72963-72758	Intron-sense
MS29_531	71846..71322	49-139	AAATGAGATAATTA AGAGTAAAATAAT GAGTAA	33	2.8	71798-71708	Intron-0654	71423..70900	74-480	TAATTAAGAGTAAAAT AAGAGTAAAAATGAGA	32	12.6	71350-70944	Intron-sense
		302-476	GTAAAAATGAGAT AATTAAGAGTGAA ATAATTA	33	5.3	71528-71371	Intron-0654		x	x	x	x	x	x
MS30_531	57386..57177	44-147	ATTTGGTGAATA	12	8.5	57343-57240	Exon_0649	57242..57033	44-171	ATTTGGTGAGTA	12	10.7	57199-57072	Exon-0649
MS31_531	51323..51047	nil	-	-	-	-	-	51181..50906	101--254	TGTAATAAA	9	16.3	51081-50928	Intron_0646

Chromosome 4: Only forward primer co-ordinates of mini-satellites MS39 and MS221/1 while reverse primer co-ordinates of MS42, MS43 and MS221/2 were located in *T. parva* Marikebuni genome.

Satellite	*T. parva* Marikebuni							*T. parva* Muguga						
	Amplicon coodinates	TRF VNTR index	VNTR pattern	VNTR size	Copy no.	VNTR Precise coodinates	Physical loci	Amplicon coodinates	TRF VNTR index	VNTR pattern	VNTR size	Copy no.	VNTR Precise coodinates	Physical loci
ms11	461658..461298	98-145	AGAGAT	6	8	461561-461514	Exonic_0241	457616..457256	98-145	AGAGAT	6	8	457519-457472	Exonic_0241
		x	x	x	x	x	x		194-304	GATAGA	6	17.5	457423-457310	Exonic
MS32	122146..122296	45-127	AACATAATT	9	9.2	122190-122272	Intron-0069	121357..121507	35-124	CAACATAAT	9	10	121391-121480	Intron-compl
		x	x	x	x	x	x		45-127	AACATAATT	9	9.2	121401-121483	Intron-compl
MS33	164918..165130	nil	-	-	-	-	-	163539..163751	35--165	ATATAGTTAATT	12	10.9	163573-163703	Intron
MS34	483915..483667	65-176	ACTATTTCCAT	11	11.5	483851-483740	Intron_0248	479007..478759	65-176	ACTATTTCCAT	11	11.5	478943-478821	Intron
MS35	510004..510163	58-121	ACTATTAAC	9	7.1	510060-510123	IGR 3'_0259	504891..505050	39-120	ACTATTAAC	9	9.1	?	IGR 3' _0259
MS36	525929..526148	32-166	TATTTATTATAC	12	11.3	525960-526094	Intron_0265	520766..520984	31-165	TATTTATTATAC	12	11.3	520796-520930	Intron
MS37	614615..614826	50-158	TATAAATTGTA	11	9.4	614664-614772	Intron-0311	609296..609496	50-147	TATAAATTGTA	11	8.4	609345-609442	Intron-compl
MS38	614845..614605	74-182	TATATACAATT	11	9.4	614772-614664	Intron_0311	609515..609286	74-171??	TATATACAATT	11	8.4	609469-609369	Intron
MS39	813112..813375	incomplete	-	-	-	-	intron	807350..807612	43-141	TTTACACA	8	12.4	807392-807490	Intron
MS40	828055..828243	87-118	GAATTAATAAATA	13	2.5	828141-828172	Intron_0417	821953..822141	47-154	GAATTAATAAATA	13	8.2	821999-822106	Intron
MS41	830362..830635	36-197	ATAGTTAATTTAC	13	12.8	830397-830658	Intron_0419	824260..824553	36-197	ATAGTTAATTTAC	13	12.8	824295-824459	Intron
		33-223	TTTACAGTTAA	11	15.5	830394-830584	Intron		33-223	TTTACAGTTAA	11	15.5	824292-824482	Intron
MS42	..832238	incomplete	-	-	-	-	-	825884..826135	51-211	ATTTATATAGTT	12	13.5	825934-826094	Intron
MS43	..832885	incomplete	-	-	-	-	-	826465..826774	45-213	AATTGTTGACT	11	15.4	826509-826677	Intron-compl
MS44	840168..840423	89-120	CAATTGAGTATA	12	2.7	840256-840287	Intron_0423	834009..834263	85-186	ATACAATTGAGT	12	8.5	834093-834194	Intron
MS45	1614530..1614951	166-366	TACACATTTT	10	19.4	1614695-1614895	Intron_0816	1606841..1607157	166-261	TACACATTTT	10	9.4	1607006-1607101	Intron
MS46	1821482..1827912	153-230	TCAACCATA	9	8.7	1821634-1821711	exon _0918	1818602..1818764	42-128	TCAACCATA	9	9.7	1818643-1818729	Exonic_0920
		153-272	TCAACCTTA	9	13.7	1821634-1821753	Exon_0918		x	x	X	x	x	X
		6283-6333	TCAACCATA	9	5.7	1827764-1827814	exon _0920		x	x	X	x	x	X
		6329-6361	CAACCTCAG	9	3.7	1827810-1827842	exon _0920		x	x	X	x	x	X
		6356-6397	CAACCTTAC	9	4.7	1827837-1827878	exon _0920		x	x	x	x	x	x
MS221/1	1643198..	incomplete						1634557..	?	?	?	?	?	?
MS221/2	..1642429	incomplete						1633982..1633813	81-142	AAGTATAGAAATAGTATAGA	20	3.2	1633902-1633841	Intron

4.2.1 *Location of VNTRs in the coding ORFs*

Some of the VNTR pattern sequences (e.g. ms8, ms10, MS12, MS19, MS817, MS221/1 and MS221/2) were aligned with other available clones (**Appendix 5**). However, the alignments produced incongruent results meaning that they were variants from the known sequences. Examination of the precise physical placement of the VNTR satellites within the genomes is shown in **Table 5**. The *T. parva* Marikebuni genome reported 10 satellites in the IGRs, 22 in the introns and 7 were in the intragenic/exonic regions. The differences in number between the two genomes are due to un-sequenced gaps in *T. parva* Marikebuni genome.

Table 5: Physical placement of VNTRs within the linear nuclear genomes of *T. parva* Marikebuni and *T. parva* Muguga.

Chromosome	*T. parva* Marikebuni			*T. parva* Muguga		
	IGR	Introns	Exons	IGR	Introns	Exons
1	ms1, ms5, MS1, MS5, MS6	ms2, ms4, MS2, MS3, MS10	ms3, MS817	ms1, ms5, MS1, MS4, MS5, MS6	ms2, ms4, MS2, MS3, MS7, MS8, MS9, MS10	ms3, MS817
2	MS11, MS19	ms6, MS17, MS20	MS13, MS14	ms7, MS11, MS16, MS19	ms6, MS12, MS15, MS17, MS18, MS20	MS13, MS14
3	MS22, MS26	MS21, MS24, MS25, MS28, MS29	MS30	ms8, MS22, MS26	ms9, MS21, MS23, MS24, MS25, MS27, MS28, MS29, MS31	MS30
4	MS35	MS32, MS34, MS36, MS37, MS38, MS40, MS41, MS44, MS45	ms11, MS46*	MS35	MS32, MS33, MS34, MS36, MS37, MS38, MS39, MS40, MS41, MS42, MS43, MS44, MS45	ms11, MS46

The majority of the VNTRs are located in the non-coding regions, especially the introns than IGRs. These results suggest a high structural conservation level within the VNTR sequences. In both genomes, only 7 VNTRs are located in the exonic regions. IGR, intergenic regions; ms - microsatellite; MS - minisatellite. Minisatellite 46 (marked *) in the *T. parva* Marikebuni genome reported a large amplicon that spanned four gene loci. Unexpectedly, the coordinates of MS221/1 and MS221/2 could not be predicted in the *T. parva* Muguga genome.

4.2.2 Protein domain orthologs of VNTR-containing ORFs

Predicted functional protein domains of the ORFs containing the VNTRs searched from interproscan database are presented in **Table 6**. The results show an even distribution of these ORF-VNTRs across the genomes with chromosome three having only one. Of the seven VNTR-containingORFs, those with MS30 and MS46, TPM02_0649 and TPM01_0920, had no significant domain hits. Two ORFs, TPM02_0615 and TPM04_0241, had two different domains each. All the domains are single copies in each of the ORFs. Only two domains have express link to the VNTR repeats i.e. thrombospondin, type 1 repeat and armadillo-type fold repeat while the remaining domains are actually in non-repetitive regions of the ORFs.

Table 6: Protein domains of the *T. parva* Marikebuni ORFs containing VNTRs retrieved from interproscan database as by 04/12/2009. All the domains are single copies in each of the ORFs. The ORFs TPM02_0615 and TPM04_0241 report two different domains each. Gene loci 0649 and 0920 containing MS30 and MS46, respectively, had no matching protein domain hit while gene 0413 bearing MS13 codes for a protein of unknown function. The entropy values (e-values) are highly significant probabilities of the correct domain hits.

Locus_tag	Satellite	Hit accession number	Short protein name	Domain hit index	E-values
TPM01_0530	ms3	IPR13170	mRNA splicing factor, Cwf21	46-89	6.40E-006
TPM01_0698	MS817	IPR001650	DNA/RNA helicase, C-terminal	696-785	2.30E-015
		IPR007502	Helicase-association region	524-918	2.00E-006
TPM02_0413	MS13	IPR013180	protein of unknown function	2-108	2.60E-020
TPM02_0615	MS14	IPR018957	Zinc finger, C3HC4 RING-type	172-212	4.20E-016
		IPR000884	Thrombospondin, type 1 repeat	848-901	9.90E-006
TPM03_0649	MS30	No hit	-	-	-
TPM04_0241	ms11	IPR003890	MIF4G-like, type 3	673-956	2.40E-022
		IPR016024	Armadillo-type fold	662-957	3.50E-056
TPM04_0920	MS46	No hit	-	-	-

4.2.3 VNTR sequence insertion point and effects on ORF

The analysis of how the VNTR sequences are inserted in protein-coding DNA regions are shown in **Table 7**. Examinations of these satellites indicate they have repeat units that are multiples of three, as shown by the 'codon size' column or else the position of insertion is altered so as to maintain the reading frame. The VNTRs are inserted at first, second and third base positions depending on their sizes to maintain the frameshifts. Only MS7 at locus 0316 in chromosome 1 shows an exception to the general trend by causing a frame-shift in the Marikebuni genome. However, this could be due to an incorrect gene model.

Table 7: Analysis of insertion of VNTRs in ORFs and their effects on the reading frames of the ORFs in terms of number of codons, base position of insertion and effect on the open reading frame. All the insertions of the VNTRs in the ORFs show no frame-shifts in reading frames. The slash (/) indicates the start and end position of repeat bases preceded by a dash (-) at the codon position.

VNTR_ORF	VNTR pattern sequence	Codon size	Insert position	Effect on Reading-frame
ms3_chr1_0530	TAGTAGAAG	Both 36	3rd base: AG-/T,AG...../TT	No frame shift
MS817_chr1_0698	ATTTTACACT	Both 25	2nd base: C--/AT, TTT.....TTA, C/CC	No frame shift
MS13_chr2_0413	TGTGTAAAA	Both 29.1	3rd base: TC_/A, CAC..../CTT	No frame shift
MS14_chr2_0615	TTTTACACAT	Mar – 20 Mug – 157	2nd base: A--/TT, ...CA-/C, A	No frame shift
MS30_chr3/531_0549	ATTTGGTGAATA	Mar – 34.1 Mug – 42.1	2nd base: AGT,C--/CA,....AAT/AGC	No frame shift
ms11_chr4_0241	AGAGAT	Both 16	1st base: TAC/AGA,.....GAT/TCG	No frame shift
	GATAGA	Mug 36.2	1st base: AGG/GAT, ...TAT/AGA	No frame shift
MS46_chr4	TCAACCATA	Mar 16.2 Mug 28.2	3rd base: AT-/T,CAA...../TA or TG	No frame shift

4.3. Codon Usage bias

The primers used in the PCR amplification and sequencing work were all the same as shown in **Appendix 2**. The sequencing results were written both in Pearson FASTA file format and ABI trace files, a section of trace file is shown in **Figure 5** below. All the sequence files, including both forward and reverse sequences for each ORF gap, were edited and assembled using Staden software (version 2.0) which also aided insertion of the consensus patterns into their respective regions in the genomes, the ORFs containing incomplete VNTRs.

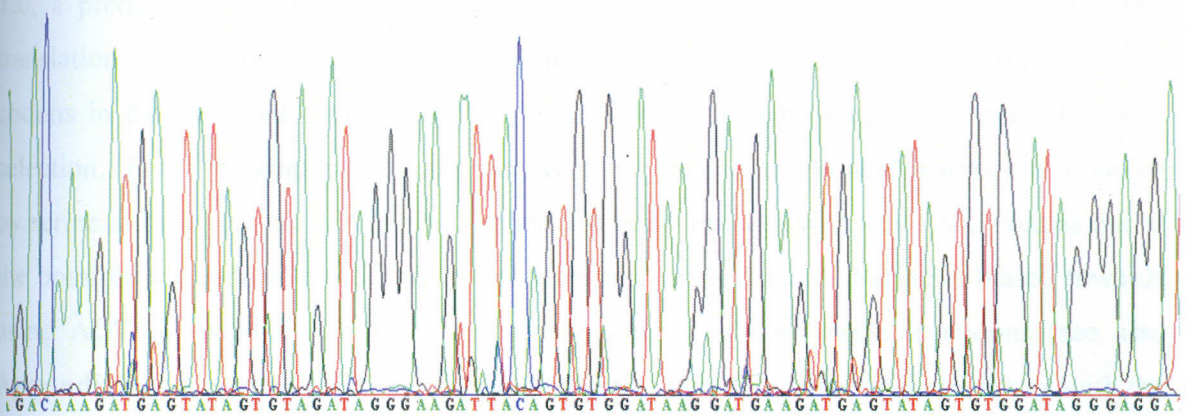

Figure 5. A section of ABI trace file showing in-part base amplitude quality of the sequenced products. The nucleotide trace amplitudes are clearly distinct and well spread out without overlaps.

4.3.1 Relative Synonymous Codon Usage (RSCU) and Relative Adaptiveness of Codons (RAC).

The complete ORFs of *T. parva* Marikebuni and their orthologs in *T. parva* Muguga were analyzed for codon bias usage in CodonOptTable program whose graphical results of RSCU versus RAC are shown in **Figures 6 - 13,** each succeeding gragh being orthologous in *T. parva* Marikebuni and *T. parva* Muguga respectively. Figures 6 (i) and (ii) has no outright highest optimal codon for locus 0920 in chromosome 4. Figure 7(i) and (ii) represent locus 0241 in chromosome with highest synonymously coded amino acids being arginine and serine. Figure 8(i) and (ii) represent locus 0649, the only coding ORF with

VNTRs in chromosome 3, synonymously coding more of arginine, proline and serine residues. In chromosome 2, represented by figures 9 and 10, the highest coded residues are alanine, arginine, and serine in locus 0413 and arginine, glycine, ileucine, leucine and serine in locus 0615 respectively in both genomes. In chromosome 1 represented in figures 11 and 12 , loci 0530 and 0698 code more of arginine, glycine, leucine, serine, and arginine, threonine, leucine respectively in both genomes. The locus CDS 0287 (coding for antigenic p67), which does not contain tandem repeats, was included because of its reported importance and as a control query sequence, is represented in figure 13(i) and (ii). In absence of codon bias usage, the RSCU values are equal to 1.0. RSCU values above this 1.0 thresh-hold give synonymous codons under selection for optimal use, while lower values indicate codons free of selection bias. All optimal codons have maximum RAC values of 1.0, a prediction that these codons are likely to be used with a lot more ease in gene translation and expression process. Of the total 59 synonymous codons, averagely, 24 codons in the analyzed data coding for 18 amino acids seem to be used optimally under selection. From the graphs, the most synonymously coded amino acid is arginine followed by serine in both genomes. Arginine is biased to using mainly AGA and AGG codons out of the possible six codons, the former codon used almost as default. Serine is biased towards using AGT followed by TCA from amongst the possible six possible codons (**see also Appendix 6)**. The trend continues for all degenerate codon species where members with the most AT-rich codons are used at higher frequencies.

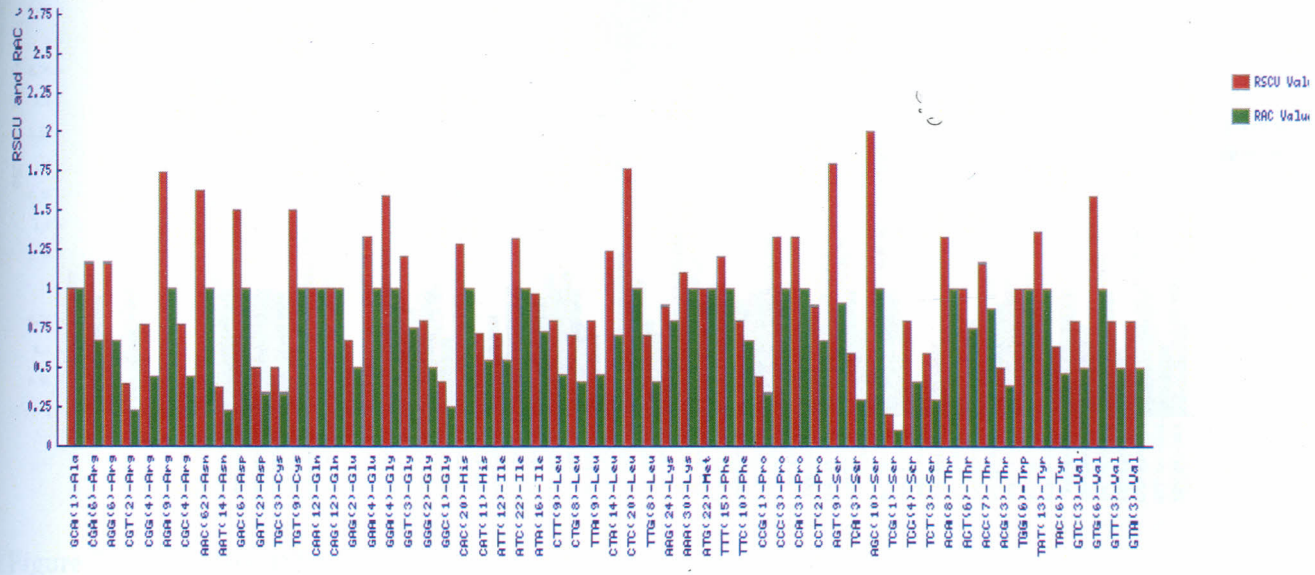

Figure 6 (i). TPM04_0920

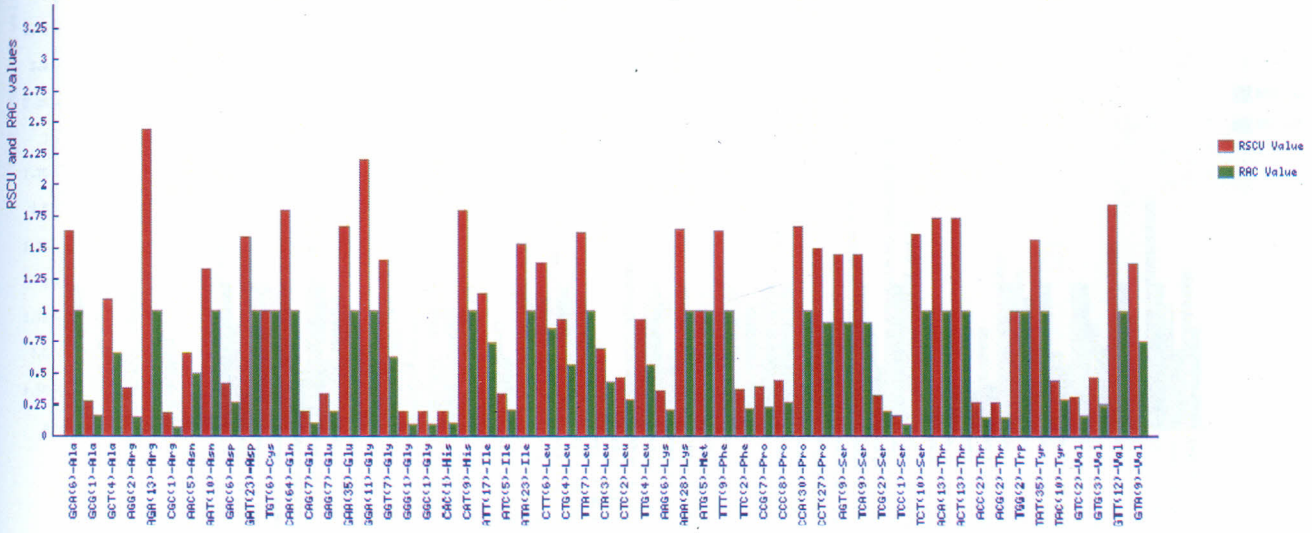

Figure 6 (ii). TP04_0920

Figure 6. CodonOptTable histograms representing Relative Synonymous Codon Uses (RSCU) and Relative Adaptiveness of Codons (RAC) of coding ORFs containing VNTRs chromosome 4 locus 0920. (i) *T. parva* Marikebuni at locus TPM04_0920 codes serine residue using AGC at highest level (tallest bar) while (ii) *T. parva* Muguga at the same locus codes arginine highest using AGA codon in preference to other synonymous codons,

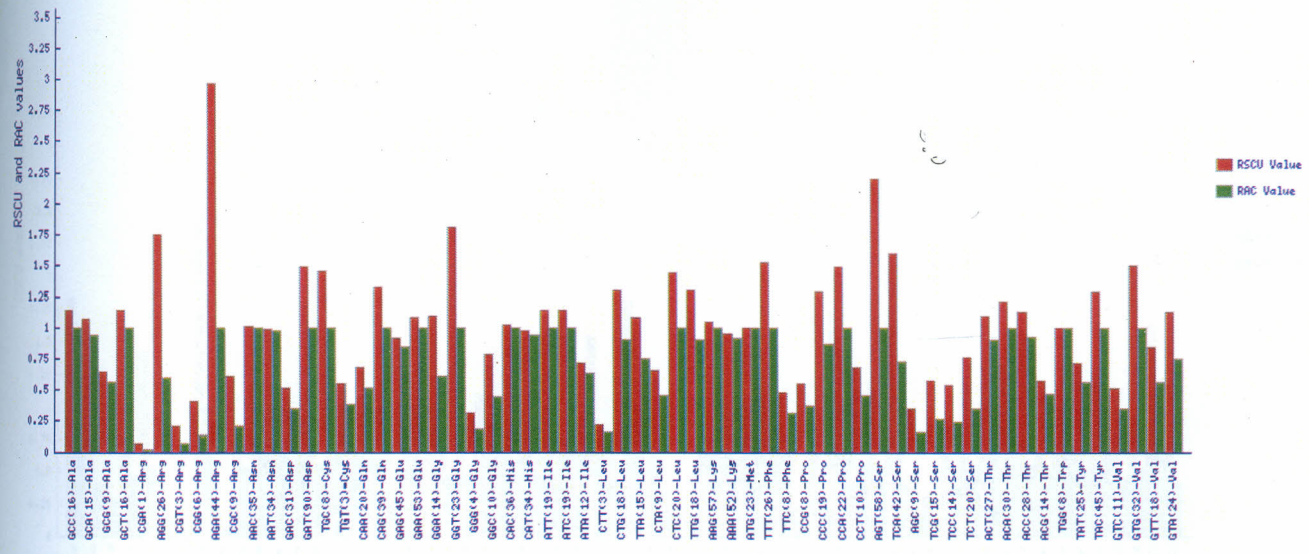

Figure 7 (i). TPM04_0241

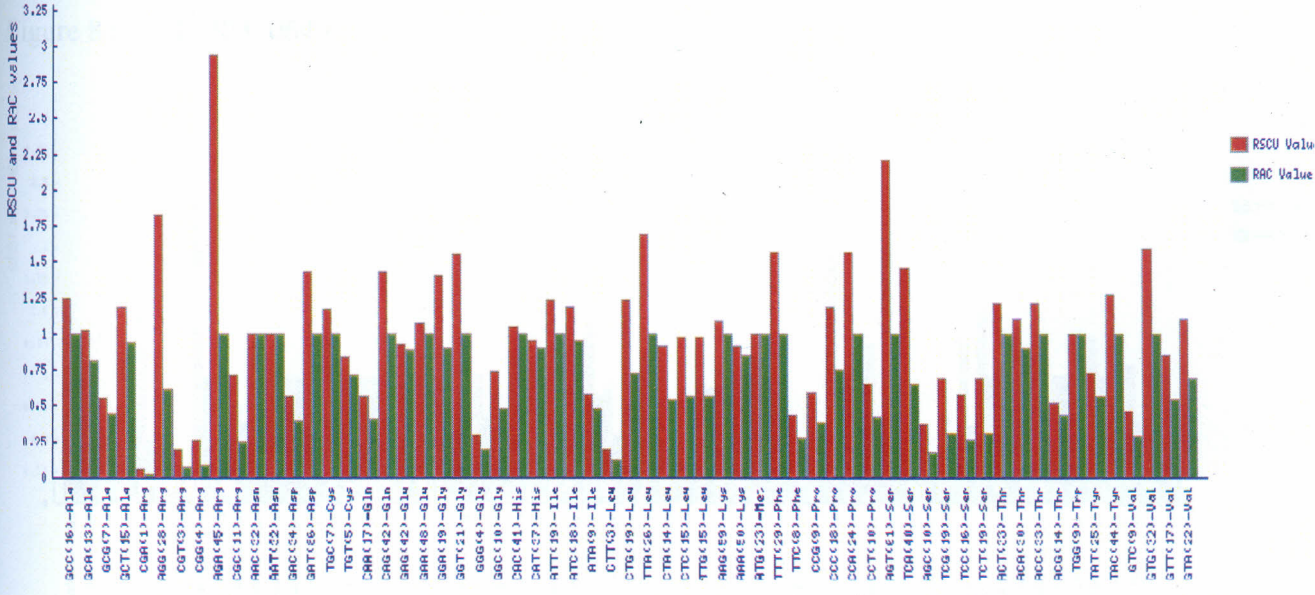

Figure 7 (ii). TP04_0241

Figure 7. CodonOptTable histograms representing Relative Synonymous Codon Uses (RSCU) and Relative Adaptiveness of Codons (RAC) of coding ORFs containing VNTRs chromosome 4 locus 0241. (i) *T. parva* Marikebuni at locus TPM04_0241 codes argine residue using AGA at similar level as in (ii) *T. parva* Muguga at the same locus followed by serine residue using AGT codon.

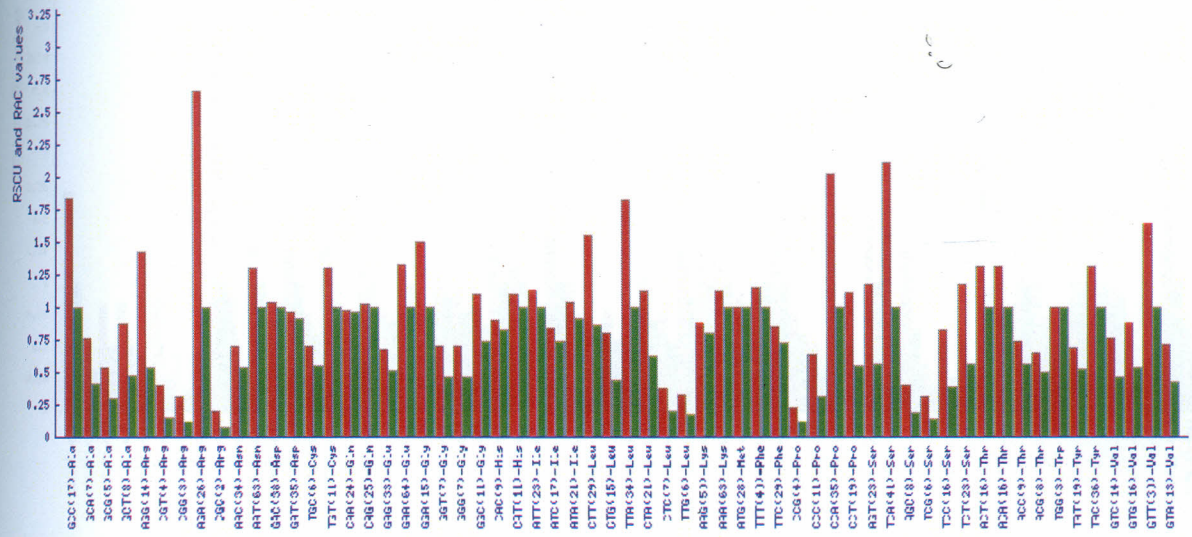

Figure 8 (i). TPM03_0649

Figure 8 (ii). TP03_0649

Figure 8. CodonOptTable histograms representing Relative Synonymous Codon Uses (RSCU) and Relative Adaptiveness of Codons (RAC) of coding ORFs containing VNTRs in chromosome 3 locus 0649. (i) *T. parva* Marikebuni at locus TPM03_0649 codes arginine, serine, proline, alanine and leucine at highest frequencies similar to (ii) in *T. parva* Muguga at the same locus,

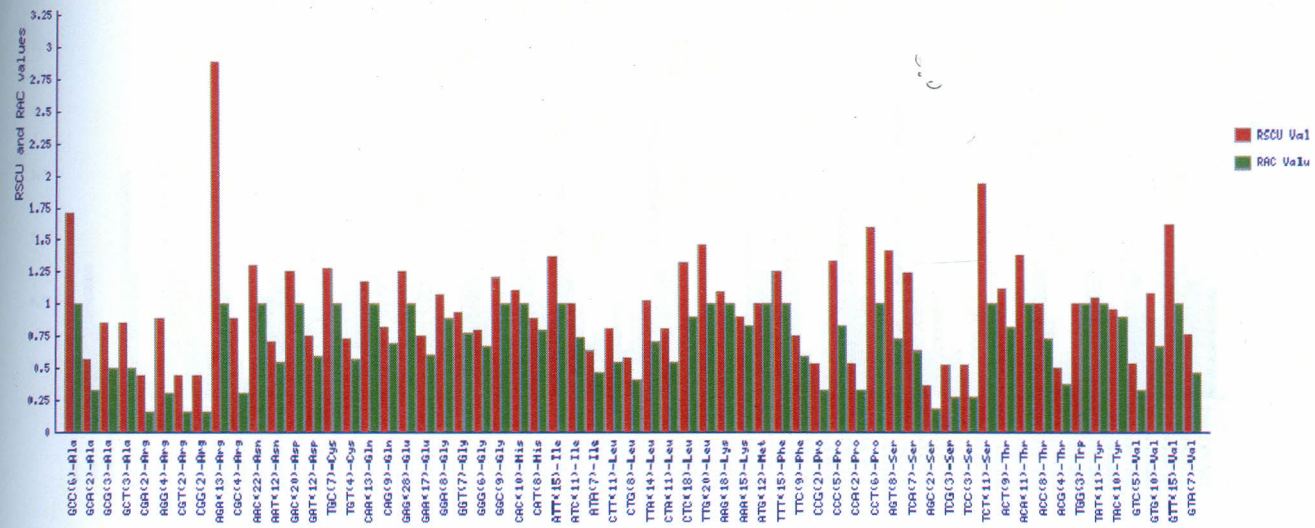

Figure 9(i). TPM02_0413

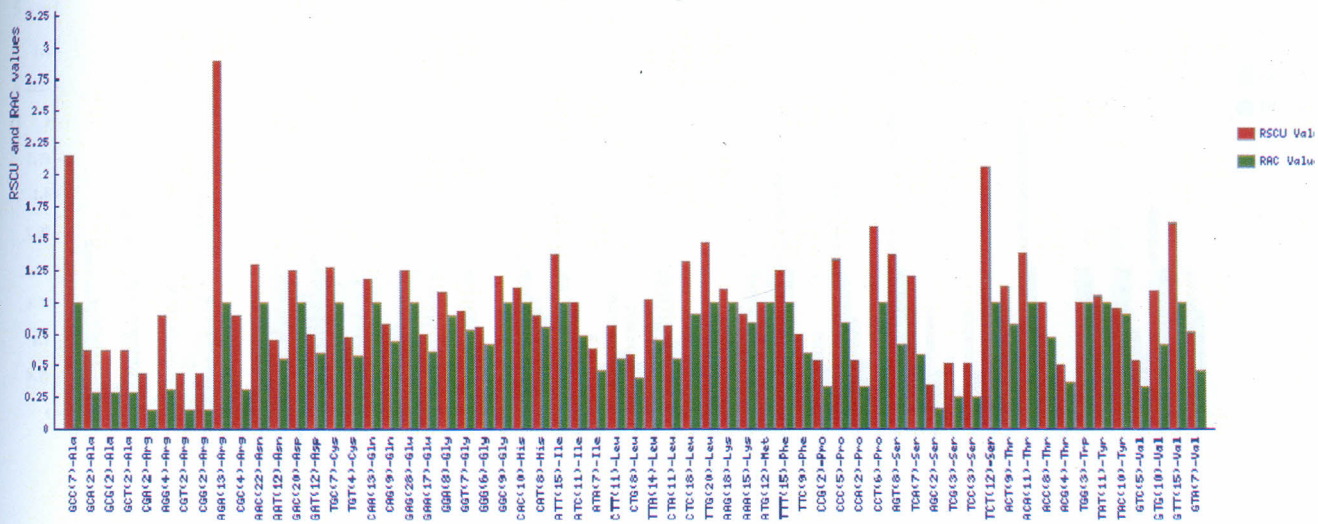

Figure 9 (ii). TP02_0413

Figure 9. CodonOptTable histograms representing Relative Synonymous Codon Uses (RSCU) and Relative Adaptiveness of Codons (RAC) of coding ORFs containing VNTRs in chromosome 2 locus 0413. (i) *T. parva* Marikebuni at locus TPM02_0413 codes more argine, serine and alanine residues at highest levels (tall bars) as the case in (ii) *T. parva* Muguga at the same locus,

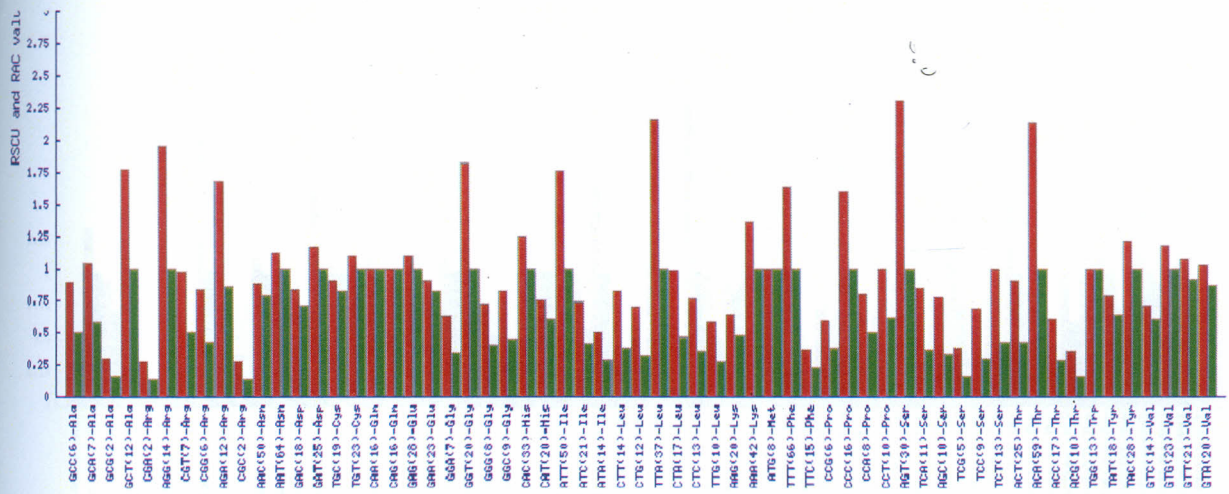

Figure 10 (i). TPM02_0615

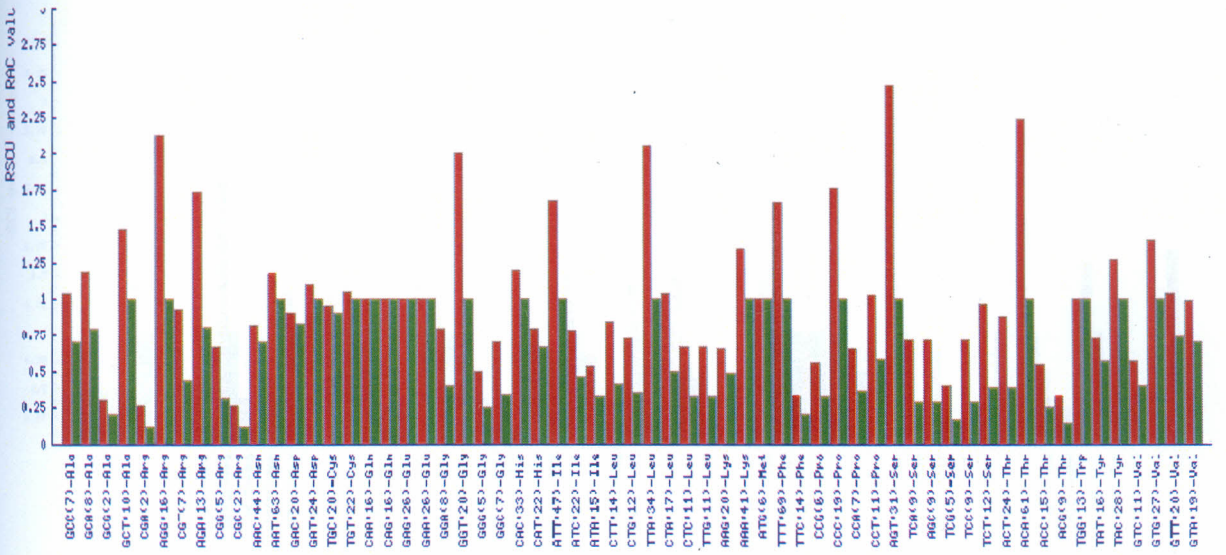

Figure 10 (ii). TP02_0615

Figure 10. CodonOptTable histograms representing Relative Synonymous Codon Uses (RSCU) and Relative Adaptiveness of Codons (RAC) of coding ORFs containing VNTRs in chromosome 2 locus 0615. In (i) *T. parva* Marikebuni at locus TPM02_0615 codes serine highly followed by leucine, threonine, argine, alanine and glycine in that order while in (ii) at the same locus, *T. parva* Muguga codes highly serine, followed by arginine, threonine, leucine, glycine among the optimal synonymous codons,

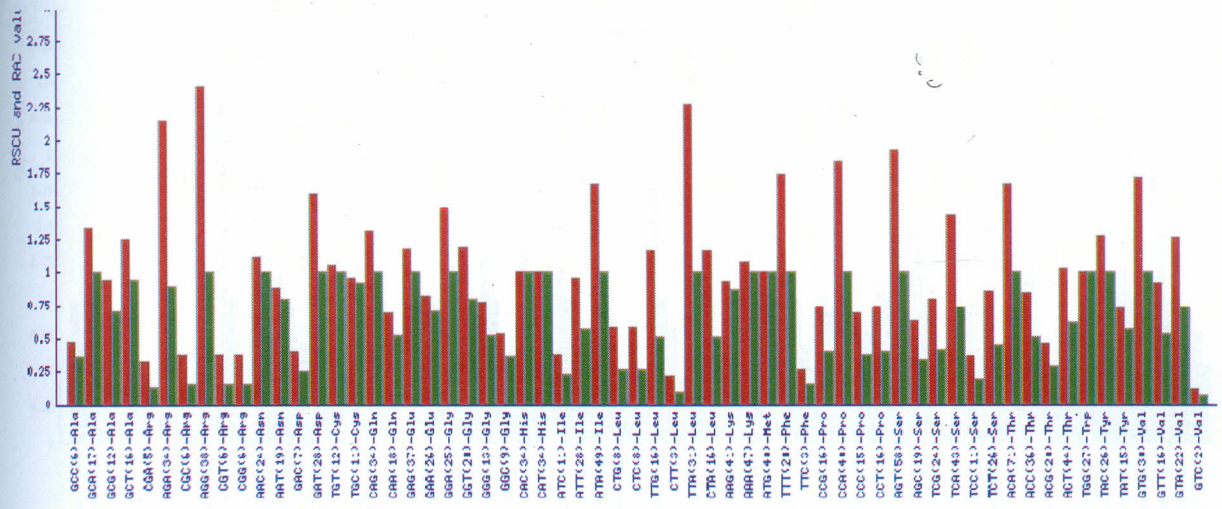

Figure 11 (i). TPM01_0530

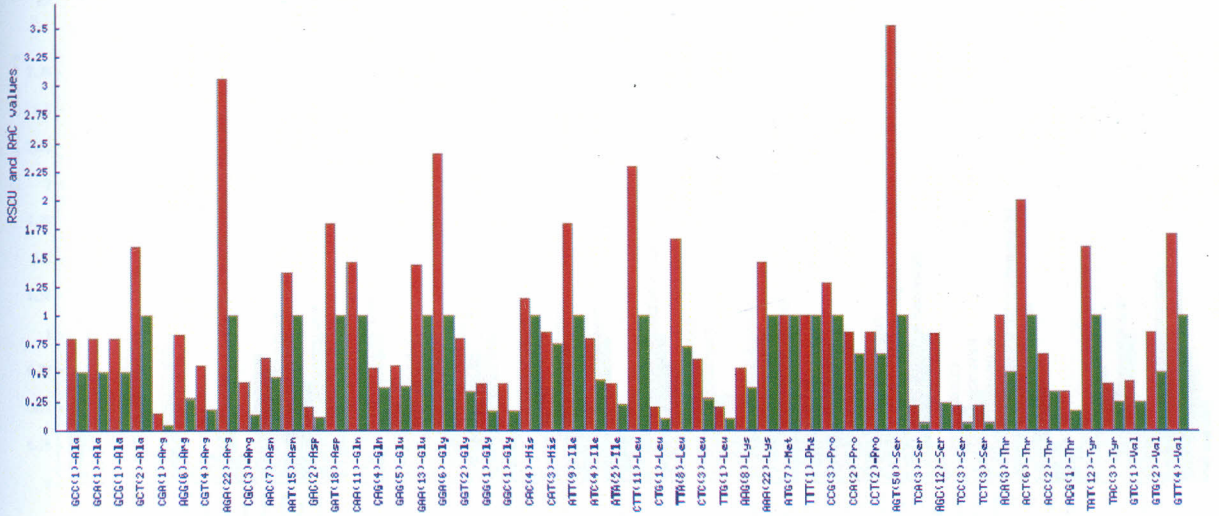

Figure 11 (ii). TP01_0530

Figure 11. CodonOptTable histograms representing Relative Synonymous Codon Uses (RSCU) and Relative Adaptiveness of Codons (RAC) of coding ORFs containing VNTRs in chromosome 1 locus 0530. The most optimmaly coded synonymous residues in *T. parva* Marikebuni at (i) locus TPM01_0530 are arginine using AGG and AGA, leucine (TTA) and serine (AGT) codon compared to serine (AGT), arginine (AGA) and glycine (GGA) as shown in (ii) TP01_0530 locus for *T. parva* Muguga.

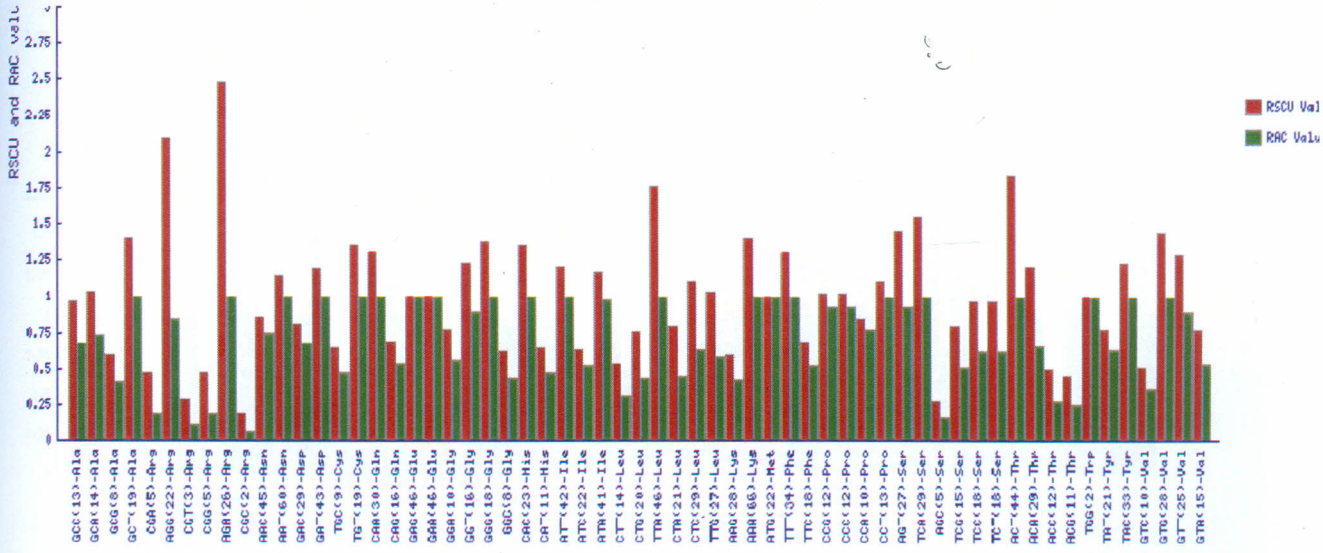

Figure 12 (i). TPM01_0698

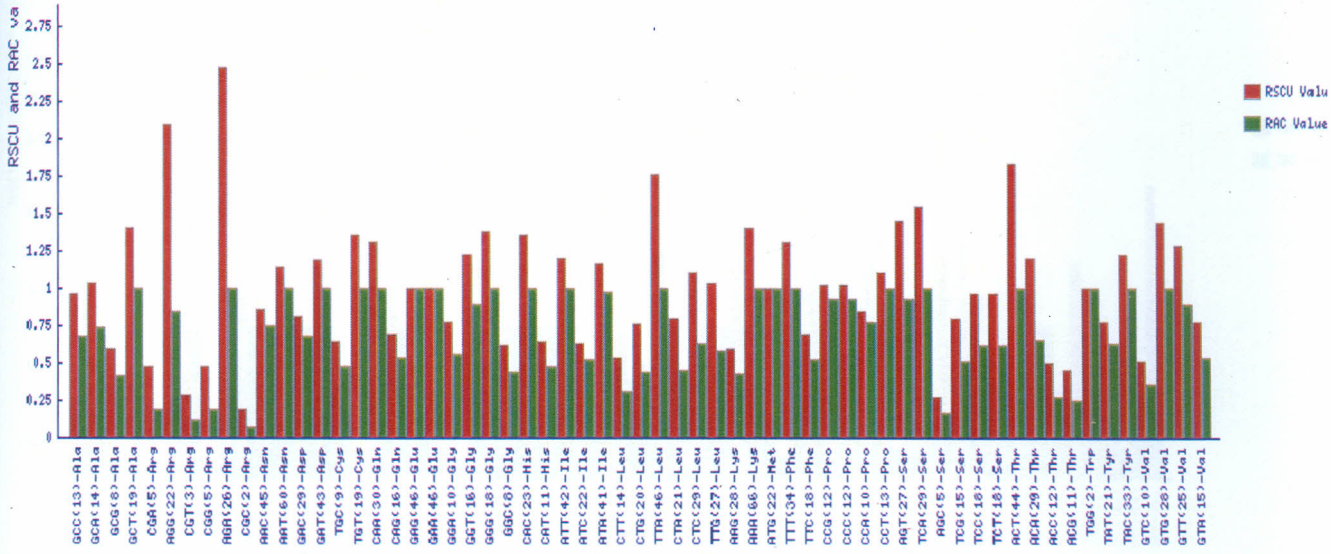

Figure 12 (ii). TP01_0698

Figure 12. CodonOptTable histograms representing Relative Synonymous Codon Uses (RSCU) and Relative Adaptiveness of Codons (RAC) of coding ORFs containing VNTRs in chromosome 1 locus 0698. In (i) locus TPM01_0698 of *T. parva* Marikebuni, and (ii) locus TP01_0698 for *T. parva* Muguga show all optimal codons used at equal levels.

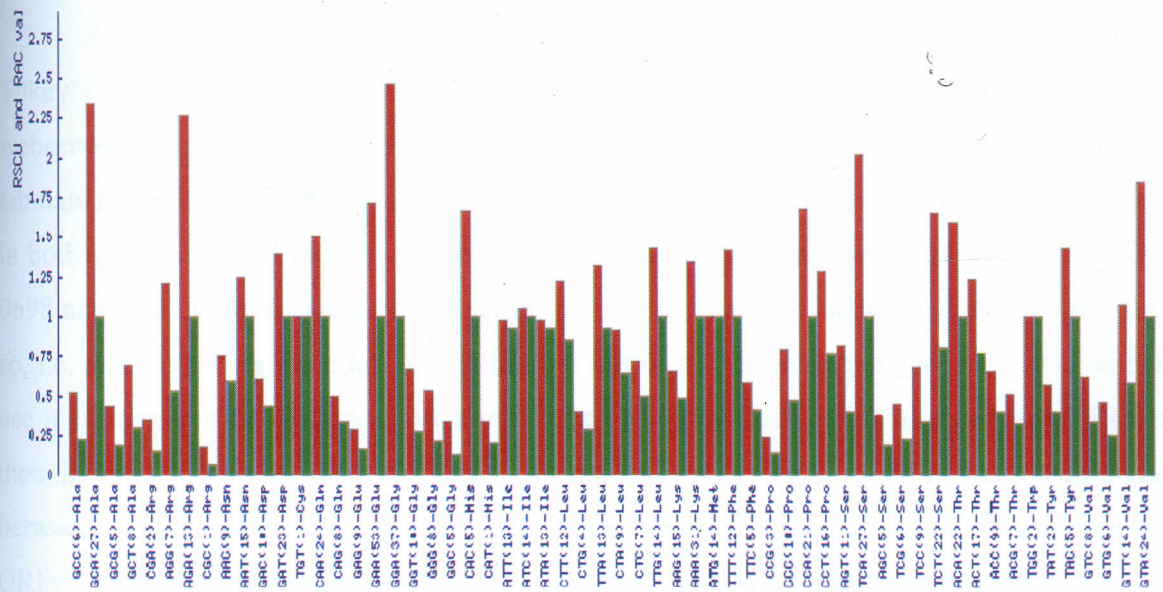

Figure 13 (i). TPM03_0287

Figure 13 (ii). TP03_0287

Figure 13. CodonOptTable histogram representing Relative Synonymous Codon Uses (RSCU) and Relative Adaptiveness of Codons (RAC) of coding ORFs containing VNTRs in chromosome 3 locus 0287. Both (i) and (ii) show equal use of the optimal codons at locus 0287 in both genomes indicating high level of conservation than the other loci.

4.3.2. Codon use statistical analysis

Synonymous codon usage bias statistical analysis within genomes was done using Tukey's Test at a significant level of 0.05. The test was implemented via CodonO webserver (http://www.sysbio.muohio.edu/CodonO/) on command line. The results are tabulated below (**Table 8**). All the ORFs indicate no significant difference in codon usage in both genomes in all the synonymous codon families. Paradoxically, the results for ORF 0698 and 0413 seem apparently significant but there is no convincing evidence to conclude so, i.e. there is a bias in codon use within the ORFs and between the genomes. The codon use measurement via Turkey's test normalizes the codon data points and therefore ensures their independence. In fact, the results do not only show similarities in both within and between genomes but also between VNTR-containing ORFs and non VNTR-containing ORFs. The p67 ORFs, which contain no VNTRs, show absence of codon usage bias.

Table 8. The statistical analysis of codon use bias within each of the two genomes was set at a significance level of $p \leq 0.05$. Locus chromosome 1_0698 and chromosome 2_0413 seem to suggest a significant difference but the evidence is not sufficient enough to support that. All the other genic regions have p-values way above 0.05, including ORFs 0698 and 0413 after rounding-up their p-values to two significant values. Therefore, the codon uses within both and between genomes are not significantly different.

Sequence identity	*T. parva* Marikebuni	*T. parva* Muguga
Chr1_0530	0.296323	0.259351
Chr1_0698	**0.0496793**	**0.0577605**
Chr2_0413	**0.0473317**	**0.0498**
Chr2_0615	0.0840489	0.0938982
Chr3_0649	0.0684748	0.0619444
Chr4_0241	0.0818594	0.0815809
Chr4_0920	0.0950399	0.277761
Chr3_0287_p67	1.0	0.153059

CHAPTER FIVE

DISCUSSION

5.1 Annotation of protein-coding DNA sequences (CDS) in *T. parva* Marikebuni genome

Theileria parva Muguga genome was the first to be sequenced and published primarily to aid in identification of schizont antigens for vaccine development and to enhance comparative genomics of related apicomplexans (Gardner *et al.*, 2005), for example, *P. falciparum* (Carlton *et al.*, 2002) and *T. annulata* (Pain *et al.*, 2005). The present study of *T. parva* Marikebuni genome annotation and analysis provides a window into the genomic characteristics of protein-coding regions of a strain thought to be more pathogenic and which will enhance comparative studies with other strains.

In concurrence with earlier reports (Guo and Silva, 2008), both genomes were evidently AT-biased. It was previously reported that *Theileria* genomes are reduced in both metabolic complexity and size relative to the genomes of other eukaryotes (Roos, 2005), and that *T. parva* Marikebuni lack some gene families altogether (Guo and Silva, 2008). The results reported here show that the *T. parva* Marikebuni genome is compact and contracted with more multi-exonic genes. The *T. parva* Marikebuni genome harbors more spliced genes as compared to Muguga strain at 3265 and 2977, respectively (however, the gene models need experimental confirmation of this observation). As previously reported, the parasite introns are indeed spread across all the reading frames in both DNA strands with recognizable consensus splice sites but with shorter lengths (Nene *et al.*, 1998; Bishop *et al.*, 2009). The introns have a visibly higher presence of stop codons with a preference for TAA codon than TAG or TGA codons. Often, this is the consequence of A-T biased genomes unlike those with high G-C contents that retain very long open reading frames (Shiels *et al.*, 2006). In support of recent reports (Chen and Manley, 2009; Koonin, 2009), the observation on multi-exons and introns could, perhaps, have a bearing on the phenotypic diversity and overall genomic complexity. Until proved, the multi-exon genes in *T. parva Marikebuni* are presumed to have conserved functionality as the orthologues in *T. parva* Muguga. Other

authors have observed that from intron-rich eukaryote genomes, in addition to conservation of coding region sequence and structure, the positions of and sequences of most introns are conserved (Carmel *et al.*, 2007; Koonin, 2009). Whether this is a case of retained ancestral introns or a recent evolution altogether remains subject to proof in *Theileria*. However, previous reports tend to suggest retention of ancestral introns, in which, the positions of many introns are known to be conserved and shared by orthologous genes even in distant eukaryotes (Carmel *et al.*, 2007). The evolutionary mechanism behind multi-exonic genes phenomenon in apicomplexans is, however, still poorly understood. Recently, it was suggested that alternative splicing and wobble-splicing could be the explanation (Lv *et al.*, 2009). In rice and Arabidopsis, the analysis for alternative splicing revealed that the majority of splice events result in a downstream frame-shift at translation, consistent with the mechanism in mouse (Haas *et al.*, 2003). The mechanism is involved in expression of nearly 95% of human multi-exon genes and in metazoans, generation of different protein products that function in diverse cellular processes, including cell growth, differentiation and death (Chen and Manley, 2009). These explanations are limited to transcript and expression level and not at DNA level and therefore cannot fit the present scenario. However, they vaguely explain the existence of higher number of multi-exonic genes in *T. parva* Marikebuni in line with its highly pathogenic and complex biology.

The presence of short non-coding regions complicates further, the understanding of the biology surrounding the binding sites of transcriptional factors. Often, transcriptional regulation is a combinatorial control involving several factors (transcriptional factors, TF) and requiring extensive specific DNA binding sites, which could be thousands of base pairs apart (Narlikar and Ovcharenko, 2009). Recently, it was suggested that the contracted *T. parva* Muguga genome could be co-transcribing a number of genes given the short non-coding regions upstream of some genes. For example, antigenic p67 have only 93 bp upstream in *T.parva* Muguga, a size doubted to be sufficient for effective binding of transcriptional factors (Bishop *et al.*, 2009). The CDSs report higher GC content in comparison to overall whole genome GC content. Taken as a pointer to *Theilerian* genome characteristic, the low GC content fits well with the contracted genome size and the presence of spliced genes. The only contrast is found in *T. mutans* and *T. orientalis,* non-transforming

species that have been reported to have higher GC content among the *Theilerias* (Nene *et al.*, 1998; Bishop *et al.*, 2009) (see **Table 2** and **Appendix 6**).

Proteins are often mosaic, containing two or more different identifiable domains, and domains can occur in different combinations in different proteins with extensive repetition of the same domain within a protein, explaining why proteins from even closely related organisms have low conservation (Rubin *et al.*, 2000). In contrast, given the relatedness of these *Theileria* clones, the intra-genic regions showed high level of structural and functional conservation as previously reported (Shiels *et al.*, 2006). Results showed a significantly lower number of the *T. parva* Muguga CDS having unrecognized domains than the *T. parva* Marikebuni genome. Most of the *T. parva* Marikebuni CDS seem to be harboring multiple paralogous functional domains fitting an earlier observation that the genome is both contracted and has high coding percentage (Pain *et al.*, 2005). This could partly be attributed to an unclear intrinsic mechanism of fine tuning of the multi-exonic CDS that could be acting at either transcriptional or translational level. Full account of genic functionality, , remains wanting given the high number of uncharacterized proteins, in line with earlier reports that *T. parva* has lots of genes whose functions are yet to be determined.

5.2 Characterization of VNTRs

The characterization of the VNTRs was based on the panel of mini-satellites and micro-satellites with their species- and locus-specific PCR primers that amplify only *T. parva* DNA as previously reported (Oura *et al.*, 2003).. Repeat sequences of short base-lengths (ranging 1- 7 bases) were defined as micro-satellites (Goldstein and Pollock, 1997) while those with longer base-lengths (8-100 bases) were defined as mini-satellites (Hamada *et al.*, 1982; Oura *et al.*, 2003). These criteria were used as classification guides in this study. Collectively, when these repeat sequences are arranged in a contiguous manner in a genome, they are usually referred to as variable number tandem repeats (VNTRs) (Oura *et al.*, 2003). Consistent with previous reports (Oura *et al.*, 2003), these *Theileria* species-specific satellites are highly conserved between the current strains of study.

5.2.1 VNTRs loci within the coding open reading frames (ORFs) of the genomes

The amplicon co-ordinates locate the primer-flanked predicted amplicons of the VNTRs in the linear chromosomal sequences of each genome while VNTR precise co-ordinates locate the actual tandem repeat sequence patterns along the linear chromosomal DNA sequences of each genome. The amplicon co-ordinates are consistent with previous chromosomal map co-ordinates of the mini- and micro-satellites (Oura *et al.*, 2003; Bishop *et al.*, 2009). The placement of VNTRs in ORFs are consistent with previous reports that VNTRs are located in both coding and non-coding DNA (Macleod *et al.*, 1999) and that they are mainly found in the non-coding regions of eukaryotic genomes (Li *et al.*, 2004a; Guo and Silva, 2008), though at a very low frequency (Nene *et al.*, 1998). However, the results presented here are restricted and remain biased to the previous reports of loci-specific VNTRs (Oura *et al.*, 2003), that is, they do not account for all the satellites that could be found in these genomes. It has been demonstrated that microsatellites are much more abundant in the untranslated regions (UTRs) or regulatory regions (for example, the transcription factors and protein kinases) and that they are non-randomly distributed across protein-coding sequences, introns and UTRs (Li *et al.*, 2004a). As such, the VNTRs in these coding regions contribute to frame shifts, fluctuation of gene expression, and inactivation of gene activity and/or change of gene function (Li *et al.*, 2004a). In quick response to environmental stress, the VNTRs located in non-translated coding regions readily aid in retuning the expression of many genes and multi-gene functions influenced by the repeat copy number in a rather general adaptive phenomena (Trifonov and Berezovsky, 2003). Essentially, VNTRs in intra-genic regions undergo a much more stringent selective pressure than those in other regions in the order of 5'-UTRs, 3'-UTRs, introns and IGRs, but this phenomenon needs further confirmation.

Examination of the few VNTR-ORFs, show that they are associated with important functional protein domains/motifs, thus explaining their high level of conservation. Analysis of the positions where the VNTRs are inserted, suggests they are under some form of constraint to avoid insertions that might change the reading frames or cause insertion of a premature terminator codon. However, it is not known how the presence of VNTRs in

ORFs affects the activities of such proteins. So far only two VNTRs have a direct association with the functional domains i.e. the thromboplastin, type 1 repeat and the armadillo-type fold. Otherwise, it can only be speculated here that they merely aid in varying the peptide lenghths. The zinc finger domains, C3HC4 RING type reportedly functional in RNA binding, mediate protein-protein interactions, and sequence-specific DNA-binding proteins and/or lipid substrates. The domains simultaneously bind ubiquitination enzymes and their substrates for degradation, hence function as ligases (http://en.wikipedia.org/wiki/Zinc_finger) and (Pabo *et al.*, 2001). The MIF4G domain is a structural motif with an ARM (armadillo) repeat type fold. The MIF4G designated type 3 is found in nuclear cap-binding proteins, eukaryotic initiation factor 4-gamma (eIF4G), and regulator of nonsense transcripts 2 (UPF2). These together with other factors promote translation initiation and progress and are crucial in nonsense-mediated mRNA decay (NMD) (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR003890). The Armadillo domain participates in cell signal transduction, regulation of desmosome assembly and cell adhesion. In *Drosophila melanongaster*, the armadillo array repeats have been associated with embryogenesis and cancerous state through *Wnt* transduction – a gene regulation via β-catenin (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR016024), but the mechanisms remain poorly understood. Thrombospondin type 1 is a neural secreted protein with multifunctional anti-angiogenic abilities. It makes use of its amino terminal which is a tryptophan-rich end to block fibroblast growth factor 2 (FGF-2) driven angiogenesis. It prevents FGF-2 binding endothelial cells, limiting their sequestration (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR000884). The DNA/RNA helicase in association with eIF4A is a diverse group of proteins that couples an ATPase activity to RNA binding and unwinding (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR001650). The mRNA splicing factor, Cwf21 participates in mRNA splicing process prior to translation (http://www.ebi.ac.uk/interpro/IEntry?ac=IPR013170).

5.2.2 Mutational mechanism and significance

Micro-satellites bear two significant characteristics of being more highly mutable and widely distributed within genomes in comparison with mini-satellites. The mutations

may be in the form of indels (insertion and deletion of one or more bases) or point mutations of one or few bases at specific loci. These mutations result in obvious changes such as copy number variation and repeat sequence pattern variants (**Appendix 1**). The known mechanisms thought to underlie these mutations are polymerase slippage in micro-satellites and gene conversion in mini-satellites (Rubinsztein *et al.*, 1995; Goldstein and Pollock, 1997). However, the occurrence and mutational mechanism of any particular VNTR affecting the operations of the genome depend on their loci. A number of models have been put forward to explain the type of mutations that occur within these satellite regions. First, the stepwise mutation model (SMM) in which alleles differ by one or few whole repeats. This is close to other author's description of mutational asymmetry i.e. a tendency to either increase (positive asymmetry) or decrease (negative asymmetry) the allele size (Petruska *et al.*, 1998). This model, however, suffers the problem of homoplasy – a case in which two satellite alleles are identical in state but not by descent (Goldstein and Pollock, 1997). Elucidating this problem in the present study is not easy because it requires more than two loci for each VNTR and a wide evolutionary gap as previously suggested (Petruska *et al.*, 1998). Second, the 'K' Alleles Model which assumes that the satellites have equal likelihood to mutate to any one K-alleles randomly (Petruska *et al.*, 1998). Third, is the Infinite Alleles Model (IAM) which assumes that each mutation can create any new allele randomly irrespective of its size (Petruska *et al.*, 1998). However, looking at the results here, it is not easy to fit the mutational mechanism into any specific model

5.3 Codon usage bias in the coding ORFs bearing VNTRs

Codons are genetic codes that transfer information encoded in nucleic acids to proteins. The codons that code for the same amino acid are referred to as synonymous codons, and they are used at relative different frequencies during translation (Grantham *et al.*, 1980; Sharp *et al.*, 1988), a phenomena referred to as codon usage bias. Codon usage analysis of synonymous class often exclude five codons – TGG and ATG that code single amino acids, Tryptophan (Trp, W) and Methionine (Met, M) respectively; and three stop

codons - UGA, UAA and UAG. In this report, only the ORFs containing the VNTRs were analyzed for codon usage bias out of curiosity of their particular location.

Codon usage bias has been found to be variable among different species including among strains and clones of a species. It is mainly related to gene function and to protein structure. The codons with high frequencies are often selectively favoured because of the ease with which they are efficiently and accurately translated, and such genes are usually highly expressed (Hershberg and Petrov, 2009). In this report, there are more than 24 codons coding for 18 amino acids, which are optimally used under selective pressure as indicated by their RSCU values above 1.0. These codons in the two genomes are, in fact, being used with no bias, meaning that their residues are highly conserved. The highest synonymously coded amino acids are arginine and serine, meaning that proteins rich in these amino acids are likely to be translated and expressed at higher levels. In support of previous reports that selection for codon usage acts in the same direction as the genomic nucleotide content (Hershberg and Petrov, 2009), the *T. parva* AT-biased genome tends to favour AT-rich codons as seen within each synonymous class (**Figure 6 and Appendix 6**). If extrapolated, these results are hoped to hold true for the whole genome analysis. In contrast to recent reports that codon usage may be different in different species even when the function of the gene is similar (Neafsey and Galagan, 2007), the statistical analyses here generally indicate a similar codon usage bias trend both within and between the genomes of study at 0.05 significance level measured relative to the overall gene GC content. As a guide in the present study, involving *T. parva* species with a higher percentage of genes having unknown functions, codon usage analysis of ORFs containing VNTRs could help in unlocking their biological function predictions.

CHAPTER SIX

SUMMARIES, CONCLUSIONS AND RECOMMENDATIONS

6.1 Summaries

6.1.1 Annotation of T. parva Marikebuni genome

- The *Theileria parva* Marikebuni nuclear genome encodes for over 3900 CDS each assigned a feature identity; majority of these CDS are multi-exonic, but are lower in number than those observed in *T. parva* Muguga. These are interspersed by smaller non-coding regions. This puts into question how the transcriptional machinery operates to increase/retain the genotypic diversity. The evolutionary mechanisms behind the generation of these multi-exonic genes in *T. parva* require an in-depth study.

- Overall, the *T. parva* Marikebuni genome is A-T rich with about 32.64% G-C content. The *T. parva* Marikebuni has a compact but a protracted genome compared to *T. parva* Muguga genome. The *T. parva* Marikebuni genome has un-sequenced (regions with unknown nucleotides) totaling about 4.12%; this is a significant amount because there could be intragenic regions yet to be sequenced.

6.1.2 Characterization of VNTRs

- This was a novel intra-specific study of VNTRs in *T. parva* strains and was able to characterize and precisely locate the 60 VNTRs within the genomes of *T. parva* Muguga and *T. parva* Marikebuni.

- VNTRs are located both in genic and non-genic regions; majority in non-genic. VNTRs have conserved repeat patterns; repeat sequence and loci in both *T. parva* genomes; but some have different repeat copy numbers.

- Insertion of VNTRs in intra-genic seem to be under selection pressure to avoid frame-shifts in ORFs; their protein domains are equally conserved.

- These results will be crucial in building *T. parva* database and make mining the satellites for any future studies quite easy , for example, in defining strain-specific markers of the *Theileria* members.

6.1.3 Codon usage bias

- The codon usage bias analysis was limited to the observed presence of the VNTRs in the intragenic regions. About 24 synonymous codons coding for 18 amino acids are optimally used under un-known selective pressure as indicated by their RSCU values above 1.0.

- Statistically there is no significant difference in the codon usage both within and between the *Theileria* genomes. This contradicts an earlier assertion that the codon usage in different species could be different even if their gene functions are the same.

6.2 Conclusions

6.2.1 Annotation of T. parva Marikebuni genome

- The *T. parva* Marikebuni nuclear genome is smaller and compact with high coding density. The CDSs have a perfect synteny to the template genome but have more multi-exonic genes. The protein functions and ontologies are also conserved but with more observed motifs and domains of unknown functions.

6.2.2 Characterization of VNTRs

- The VNTRs in both genomes are similar in terms of repeat patterns and repeat sequences except in repeat copy numbers. The results indicated that more VNTRs are located in the non-coding regions of DNA, the intergenic regions and introns than are found in the protein-coding ORFs of both genomes.

- These results will be crucial in building *T. parva* database and make mining the satellites for any future studies quite easy, for example, in defining strain-specific markers of the *Theileria* members.

6.2.3 Codon usage bias

- The codon usages in these genomes are biased to using AT-rich codons as would positively be expected of AT-rich genomes.
- The presence of VNTRs in the genic ORFs does not affect the codon usage, which remains conserved between the two *Theileria parva* genomes.

6.3 Recommendations

6.3.1 Recommendations for applications

- The findings on CDS annotation in *T. parva* Marikebuni has increased the number of annotated *T. parva* species genomes and are currently being used to enhance intra-specific genomic comparative studies of the other *T. parva* strains, including *T. parva* Katete, *T. parva* Serengeti, and *T. parva* Muguga-Marikebuni recombinant.
- The results on CDS feature identification tags and VNTR genomic coordinates have been used in fast-tracking a *T. parva* database (Tpdbase) to ease data mining; specifically, with just a query line, for example, a CDS locus tag or a VNTR code (TPM02_0215 or MS37, respectively), the exact co-ordinates are located with a filter on chromosome number.
- With the full list of CDS protein functions and ontologies in both *T. parva* Marikebuni and *T. parva* Muguga genomes, search for the functionality of putative and unknown proteins will be easy, for instance, via a 3D structure pipeline like a PSPP (protein structure prediction pipeline) package (Lee *et al.*, 2009).
- The bioinformatic approaches used in this study has proved robust in predicting the exact loci of the CDSs in the genomes. All the CDSs can now be used to carry out a full expression profiling to help show that they are true gene models.

- The results showing presence of multi-exonic genes within a smaller genome is a good pointer towards elucidating genotypic diversity of *T. parva* Marikebuni, especially when comparative studies with other clones/strains are completed.

- The analysis of codon usage indicates no bias resulting from the presence of VNTRs in the protein-coding regions of the genomes. This improves the understanding of previous hypotheses that codon usage could differ between orthologous genes of organisms of the same species.

6.3.2 Recommendations for further research

- The conclusions drawn in this study are based on a partial nuclear genome of *T. parva* Marikebuni, and the structural annotation of the CDS will need a review once the genome re-sequencing which is currently on-going is completed.

- There is a need to perform an experimental validation to confirm that the gene models adopted here are correct and that the multi-exonic splice points are not merely artefacts but genuine spliceomes of the genes. This should include a study of the mechanisms behind the generation of these multi-exons and their possible effect on the genotypic diversity of the Marikebuni genome.

- All the CDS proteins with unknown functions and putative functions should be subjected to protein folding search such as PSPP so as to identify if there are novel proteins, pseudogenes, or just gene segments.

- The effects of VNTRs located in the ORFs on the functioning of the coded proteins should be investigated further.

- There is need to perform a whole genome codon usage analysis in at least one member of *Theileria* whose annotations have been completed to act as a reference genome for any future comparative study.

REFERENCES

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

Beck, R., Vojta, L., Mrljak, V., Marinculic, A., Beck, A., Zivicnjak, T. and Caccio, S.M. (2009). Diversity of *Babesia* and *Theileria* species in symptomatic and asymptomatic dogs in Croatia. *Int. J. Parasitol.*, **39:** 843-848

Behnke, M.S., Radke, J.B., Smith, A.T., Sullivan, W.J., Jr. and White, M.W. (2008). The transcription of bradyzoite genes in *Toxoplasma gondii* is controlled by autonomous promoter elements. *Mol. Microbiol.*, **68**: 1502-18.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**: 573-80.

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

Bishop, R., Morzaria, S. and Gobright, E. (1998). Linkage of two distinct AT-rich minisatellites at multiple loci in the genome of *Theileria parva*. *Gene*, **216:** 245-54.

Bishop, R., Odongo, D.O., Mann, D.J., Pearson, T.W., Sugimoto, C., Haines, L.R., Glass, E., Jensen, K., Seitzer, U., Ahmed, J.S., Graham, S.P. and E., d.V.P. (2009). Genome Mapping and Genomics in Animal-Associated Microbes. In V. Nene and C. Kole (eds.), Springer-Verlag Berlin Heidelberg 192-231.

Brayton, K.A., Lau, A.O., Herndon, D.R., Hannick, L., Kappmeyer, L.S., Berens, S.J., Bidwell, S.L., Brown, W.C., Crabtree, J., Fadrosh, D., Feldblum, T., Forberger, H.A., Haas, B.J., Howell, J.M., Khouri, H., Koo, H., Mann, D.J., Norimine, J., Paulsen, I.T., Radune, D., Ren, Q., Smith, R.K., Jr., Suarez, C.E., White, O., Wortman, J.R., Knowles, D.P., Jr., McElwain, T.F. and Nene, V.M. (2007). Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. *PLoS Pathog.*, **3**: 1401-13.

Burke, D.T. (1991). The role of yeast artificial chromosome clones in generating genome maps. *Curr. Opin. Genet. Dev.*, **1**: 69-74.

Carlton, J.M., Angiuoli, S.V., Suh, B.B., Kooij, T.W., Pertea, M., Silva, J.C., Ermolaeva, M.D., Allen, J.E., Selengut, J.D., Koo, H.L., Peterson, J.D., Pop, M., Kosack, D.S., Shumway, M.F., Bidwell, S.L., Shallom, S.J., van Aken, S.E., Riedmuller, S.B., Feldblyum, T.V., Cho, J.K., Quackenbush, J., Sedegah, M., Shoaibi, A., Cummings, L.M., Florens, L., Yates, J.R., Raine, J.D., Sinden, R.E., Harris, M.A., Cunningham, D.A., Preiser, P.R., Bergman, L.W., Vaidya, A.B., van Lin, L.H., Janse, C.J., Waters, A.P., Smith, H.O., White, O.R., Salzberg, S.L., Venter, J.C., Fraser, C.M., Hoffman, S.L., Gardner, M.J. and Carucci, D.J. (2002). Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. *Nature*, **419**: 512-9.

Carmel, L., Wolf, Y.I., Rogozin, I.B. and Koonin, E.V. (2007). Three distinct modes of intron dynamics in the evolution of eukaryotes. *Genome Res.*, **17**: 1034-44.

Carver, T., Berriman, M., Tivey, A., Patel, C., Bohme, U., Barrell, B.G., Parkhill, J. and Rajandream, M.A. (2008). Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics*, **24**: 2672-6.

Chan, C.X., Beiko, R.G., Darling, A.E. and Ragan, M.A. (2009). Lateral transfer of genes and gene fragments in prokaryotes. *Genome Biol. Evol.*, **1**: 429-38.

Chen, M. and Manley, J.L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat. Rev. Mol. Cell Biol.*, **10**: 741-54.

Delcher, A.L., Kasif, S., Fleischmann, R.D., Peterson, J., White, O. and Salzberg, S.L. (1999). Alignment of whole genomes. *Nucleic Acids Res.*, **27**: 2369-76.

Dobbelaere, D.A. and Kuenzi, P. (2004). The strategies of the *Theileria* parasite: a new twist in host-pathogen interactions. *Curr. Opin. Immunol.*, **16**: 524-30.

Dobbelaere, D.A.E. and McKeever, D.J. (2002). *Theileria*, Kluwer Academic Publishers.

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. and Nene, V. (2005). Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science*, **309**: 134-7.

Goddeeris, B.M., Morrison, W.I., Toye, P.G. and Bishop, R. (1990). Strain specificity of bovine *Theileria parva*-specific cytotoxic T cells is determined by the phenotype of the restricting class I MHC. *Immunology*, **69**: 38-44.

Goldstein, D.B. and Pollock, D.D. (1997). Launching microsatellites: a review of mutation processes and methods of phylogenetic interference. *J. Hered.*, **88**: 335-42.

Grantham, R., Gautier, C. and Gouy, M. (1980). Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res.*, **8**: 1893-912.

Grocock R.J. and Sharp P.M. (2001). Synonymous codon usage in *Cryptosporidium parvum:* identification of two distinct trends among genes. *Int. J Parasitol.*, **31:** 402-412.

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

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S.L. and White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, **31**: 5654-66.

Hamada, H., Petrino, M.G. and Kakunaga, T. (1982). A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. *Proc. Natl. Acad. Sci. U. S. A.*, **79**: 6465-9.

Hershberg, R. and Petrov, D.A. (2009). General rules for optimal codon choice. *PLoS Genet.*, **5**: e1000556.

http://blast.ncbi.nlm.nih.gov/Blast.cgi 01.10.2009

http://www.ebi.ac.uk/Tools/fasta33/index.html 01.10.2006.

http://www.ebi.ac.uk 01.10.2009

http://en.wikipedia.org/wiki/Tandem_repeat 20.11.2009

http://www.web-books.com/MoBio/Free/Ch3G1.htm 20.11.2010

http://www.tigr.org/tbd/e2k1/tpa1 12.02.2009

http://www.hpc.ilri.cgiar.org/tools 05.10.2009

http://www.tigr.org/tbd/e2k1/tpa1 12.02.2009

http://www.ncbi.nlm.nih.gov/BLAST/ 12.02.2009

http://www.ebi.ac.uk/tools/ 15.10.2009

http://hpc.ilri.cgiar.org/tools/emboss/ 20.11.2009

http://us.bioneer.com/ 20.11.2009

http://francois.schweisguth.free.fr/protocols/QIAquick_PCR_Purification_Kit 20.11.2009

http://staden.sourceforge.net/overview.html 15.12.2009

http://search.cpan.org/~shardiwal/Bio-Tools-CodonOptTable-0.07/lib/Bio/Tools/CodonOptTable.pm 05.01.2010

http://www.sysbio.muohio.edu/CodonO/ 05.01.2010

http://en.wikipedia.org/wiki/Zinc_finger 10.12.2010

http://www.ebi.ac.uk/interpro/IEntry?ac=IPR003890 10.12.2010

http://www.ebi.ac.uk/interpro/IEntry?ac=IPR016024 10.12.2010

http://www.ebi.ac.uk/interpro/IEntry?ac=IPR000884 10.12.2010

http://www.ebi.ac.uk/interpro/IEntry?ac=IPR001650 10.12.2010

http://www.ebi.ac.uk/interpro/IEntry?ac=IPR013170 10.12.2010

Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2:** 13-34.

Jura, W.G.Z.O. (1984). Factors affecting the capacity of *Theileria annulata* sporozoites to invade bovine peripheral blood lymphocytes. *Vet. Parasitol.*, **16**: 215-23.

Jura, W.G.Z.O., Brown, C.G.D. and Rowlands, A.C. (1983). Ultrastructural characteristics of in vitro parasite-lymphocyte behavior in invasions with *Theileria annulata* and *Theileria parva*. *Vet. Parasit.*, **12**: 115-134.

Katzer, F., Ngugi, D., Oura, C., Bishop, R.P., Taracha, E.L., Walker, A.R. and McKeever, D.J. (2006). Extensive genotypic diversity in a recombining population of the apicomplexan parasite *Theileria parva*. *Infect. Immun.*, **74**: 5456-64.

Koonin, E.V. (2009). Darwinian evolution in the light of genomics. *Nucleic Acids Res.*, **37:**, 1011-34.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C. and Salzberg, S.L. (2004). Versatile and open software for comparing large genomes. *Genome Biol.,* **5**: R12.

Lee, M.S., Bondugula, R., Desai, V., Zavaljevski, N., Yeh, I.C., Wallqvist, A. and Reifman, J. (2009). PSPP: a protein structure prediction pipeline for computing clusters. *PLoS One,* **4**: e6254.

Li, B., Xia, Q., Lu, C., Zhou, Z. and Xiang, Z. (2004a). Analysis on frequency and density of microsatellites in coding sequences of several eukaryotic genomes. *Genom. Prot. Bioinfo.,* **2**: 24-31.

Li, Y.C., Korol, A.B., Fahima, T. and Nevo, E. (2004b). Microsatellites within genes: structure, function, and evolution. *Mol. Biol. Evol.,* **21**: 991-1007.

Lister, R., Gregory, B.D. and Ecker, J.R. (2009). Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Curr. Opin. Plant Biol.,* **12**: 107-18.

Lv, J., Yang, Y., Yin, H., Chu, F., Wang, H., Zhang, W., Zhang, Y. and Jin, Y. (2009). Molecular determinants and evolutionary dynamics of wobble splicing. *Mol. Biol. Evol.,* **26**: 1081-92.

Macleod, D., Clark, V.H. and Bird, A. (1999). Absence of genome-wide changes in DNA methylation during development of the zebrafish. *Nat. Genet.,* **23**: 139-40.

Mardis, E.R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet.*, **24**: 133-41.

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

Narlikar, L. and Ovcharenko, I. (2009). Identifying regulatory elements in eukaryotic genomes. *Brief. Funct. Genomic Proteomic,* **8**: 215-30.

Neafsey, D.E. and Galagan, J.E. (2007). Positive selection for unpreferred codon usage in eukaryotic genomes. *BMC Evol. Biol.,* **7**: 119.

Nene, V., Morzaria, S. and Bishop, R. (1998). Organisation and informational content of the *Theileria parva* genome. *Mol. Biochem. Parasitol.*, **95**: 1-8.

Norval, R.A.I., Perry, B.D. and Young, A.S. (1992). The epidemiology of theileriosis in Africa, Academic Press, London, England.

Odongo, D.O., Oura, C.A., Spooner, P.R., Kiara, H., Mburu, D., Hanotte, O.H. and 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.*, **36**: 937-46.

Oura, C.A., Asiimwe, B.B., Weir, W., Lubega, G.W. and Tait, A. (2005). Population genetic analysis and sub-structuring of *Theileria parva* in Uganda. *Mol. Biochem. Parasitol.*, **140**: 229-39.

Oura, C.A., Odongo, D.O., Lubega, G.W., Spooner, P.R., Tait, A. and Bishop, R.P. (2003). A panel of microsatellite and minisatellite markers for the characterisation of field isolates of *Theileria parva*. *Int. J. Parasitol.*, **33**: 1641-53.

Pabo, C.O., Peisach, E. and Grant, R.A. (2001). Design and selection of novel Cys2His2 zinc finger proteins. *Annu. Rev. Biochem.*, **70**, 313-40.

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. and Hall, N. (2005). Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. *Science*, **309**: 131-3.

Petruska, J., Hartenstine, M.J. and Goodman, M.F. (1998). Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease. *J. Biol. Chem.*, **273:** 5204-10.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Brief. Bioinform.*, **10**: 354-66.

Pop, M. and Kosack, D. (2004). Using the TIGR assembler in shotgun sequencing projects. *Methods Mol. Biol.*, **255**: 279-94.

Pop, M. and Salzberg, S.L. (2008). Bioinformatics challenges of new sequencing technology. *Trends Genet.*, **24**: 142-9.

Roos, D.S. (2005). Genetics. Themes and variations in apicomplexan parasite biology. *Science,* **309**: 72-3.

Rubin, G.M., Yandell, M.D., Wortman, J.R., Gabor Miklos, G.L., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., Cherry, J.M., Henikoff, S., Skupski, M.P., Misra, S., Ashburner, M., Birney, E., Boguski, M.S., Brody, T., Brokstein, P., Celniker, S.E., Chervitz, S.A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R.F., Gelbart, W.M., George, R.A., Goldstein, L.S., Gong, F., Guan, P., Harris, N.L., Hay, B.A., Hoskins, R.A., Li, J., Li, Z., Hynes, R.O., Jones, S.J., Kuehl, P.M., Lemaitre, B., Littleton, J.T., Morrison, D.K., Mungall, C., O'Farrell, P.H., Pickeral, O.K., Shue, C., Vosshall, L.B., Zhang, J., Zhao, Q., Zheng, X.H. and Lewis, S. (2000). Comparative genomics of the eukaryotes. *Science*, **287**: 2204-15.

Rubinsztein, D.C., Amos, W., Leggo, J., Goodburn, S., Jain, S., Li, S.H., Margolis, R.L., Ross, C.A. and Ferguson-Smith, M.A. (1995). Microsatellite evolution--evidence for directionality and variation in rate between species. *Nat. Genet.,* **10**: 337-43.

Shah, T., de Villiers, E., Nene, V., Hass, B., Taracha, E., Gardner, M.J., Sansom, C., Pelle, R. and Bishop, R. (2006). Using the transcriptome to annotate the genome revisited: application of massively parallel signature sequencing (MPSS). *Gene*, **366**: 104-8.

Sharp, P.M., Cowe, E., Higgins, D.G., Shields, D.C., Wolfe, K.H. and Wright, F. (1988). Codon usage patterns in *Escherichia coli*, *Bacillus subtilis, Saccharomyces cerevisiae*, *Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. *Nucleic Acids Res.*, **16**: 8207-11.

Sharp, P.M. and Li, W.H. (1987). The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.,* **15**: 1281-95.

Shaw, M.K. (2003). Cell invasion by *Theileria* sporozoites. *Trends Parasitol.*, **19**: 2-6.

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

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**: 1135-45.

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

Shirley, M.W. and Harvey, D.A. (2000). A genetic linkage map of the apicomplexan protozoan parasite Eimeria tenella. *Genome Res.*, **10:** 1587-93.

Sohanpal, B.K., Morzaria, S.P., Gobright, E.I. and Bishop, R.P. (1995). Characterisation of the telomeres at opposite ends of a 3 Mb *Theileria parva* chromosome. *Nucleic Acids Res.*, **23**: 1942-7.

Souza, R.T., Santos, M.R., Lima, F.M., El-Sayed, N.M., Myler, P.J., Ruiz, J.C. and da Silveira, J.F. (2007). New *Trypanosoma cruzi* repeated element that shows site specificity for insertion. *Eukaryot. Cell,* **6:** 1228-38.

Sunil, S., Chauhan, V.S. and Malhotra, P. (2008). Distinct and stage specific nuclear factors regulate the expression of falcipains, *Plasmodium falciparum* cysteine proteases. *BMC Mol. Biol.*, **9:** 47.

Thompson, J., Janse, C.J. and Waters, A.P. (2001). Comparative genomics in *Plasmodium*: a tool for the identification of genes and functional analysis. *Mol. Biochem. Parasitol.*, **118**: 147-54.

Trifonov, E.N. and Berezovsky, I.N. (2003). Evolutionary aspects of protein structure and folding. *Curr. Opin. Struct. Biol.*, **13**: 110-4.

van Noort, V. and Huynen, M.A. (2006). Combinatorial gene regulation in *Plasmodium falciparum*. *Trends Genet.*, **22**: 73-8.

Vogt, P. (1990). Potential genetic functions of tandem repeated DNA sequence blocks in the human genome are based on a highly conserved "chromatin folding code". *Hum. Genet.*, **84**: 301-36.

Weir, W., Ben-Miled, L., Karagenc, T., Katzer, F., Darghouth, M., Shiels, B. and Tait, A. (2007). Genetic exchange and sub-structuring in *Theileria annulata* populations. *Mol. Biochem. Parasitol.*, **154**: 170-80.

Westesson, O. and Holmes, I. (2009). Accurate detection of recombinant breakpoints in whole-genome alignments. *PLoS Comput. Biol.*, **5**: e1000318.

Wilkowsky, S.E., Moretta, R., Mosqueda, J., Gil, G., Echaide, I., Lia, V., Falcon, A., Florin Christensen, M. and Farber, M. (2009). A new set of molecular markers for the genotyping of *Babesia bovis* isolates. *Vet. Parasitol.*, **161**: 9-18.