UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

High Resolution Analysis Of Genes Transcribed In Ixodid Tick Tissues With Special Reference To Salivary Glands Of The Brown Ear Tick, *Rhipicephalus appendiculatus.*

By

Sonal Prabhulal Henson

Submitted in partial fulfillment of the requirements for the degree

Philosophiae Doctor

In the Faculty of Natural and Agricultural Sciences
University of Pretoria
Pretoria

April 2013

© University of Pretoria

Declaration

I, Sonal Prabhulal Henson, declare that this thesis, which I hereby submit for the degree of *Philosophiae Doctor* at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institute.

Signature: ……..

Date: ……

© University of Pretoria

Acknowledgments

I thank my supervisors Drs Richard Bishop and Etienne de Villiers, for their guidance and insights into this study. I am also extremely grateful to my husband Dave for his encouragement and support while writing up this thesis.

© University of Pretoria

Table of Contents

© University of Pretoria

© University of Pretoria

© University of Pretoria

List of Abbreviations

aa	Amino acid
AvGI	*Amblyomma variegatum* Gene Index
BAC	Bacterial Artificial Clone
BmiGI	*Rhipicephalus* (*Boophilus*) *microplus* Gene Index
bp	base pairs
CDD	Conserved Domain Database
cDNA	complementary DNA
ECF	East Coast Fever
EST	Expressed Sequence Tags
Gb	Giga bases
IsGI	*Ixodes scapularis* Gene Index
Kb	Kilo bases
Mb	Mega bases
ncRNA	non-coding RNA
NR	Non-redundant protein database
NT	Non-redundant nucleotide database
RA	*Rhipicephalus appendiculatus*
RaGI	*Rhipicephalus appendiculatus* Gene Index
SSH	Suppression Subtractive Hybridization
SVM	Support Vector Machine
TC	Tentative Consensus
TIGR	The Institute of Genome Research

© University of Pretoria

List of Figures

© University of Pretoria

© University of Pretoria

List of Tables

© University of Pretoria

© University of Pretoria

Summary

R. appendiculatus, also known as the brown-ear tick, is the primary vector for *T. parva,* a protozoan parasite that causes East Coast Fever, a fatal lympho-proliferative disease of domestic cattle and of great economic importance in sub-saharan Africa. An EST dataset comprising 18,422 primary sequences was generated from the salivary gland tissues of the tick *Rhipicephalus appendiculatus* by the Institute for Genomic Research (TIGR) and selected outputs from a preliminary automated annotation were published (Nene et al., 2004). Stimulated by existence of this dataset and the nature of the initial analysis, I hypothesized regarding *R. appendiculatus* tick EST and genomic sequences as follows:

Chapter 3

That technical issues in both the quality and quantity of primary data and the analysis methods created significant biases in the composition of the currently available *R. appendiculatus* tick EST database (named RaGI):

(A) because the library used to generate the data was size selected (>2 kb) it was not fully representative and missed shorter transcripts potentially encoding functionally important peptides

(B) only potential protein coding genes were analyzed according to strictly defined database matches. No other algorithms were used, for example searching for conserved domain motifs and predicted structural features

(C) due to the redundancy of expressed tick sequences and inter-individual heterogeneity of tick sequence expression. RaGI comprising 18,422 primary sequences, although derived from a single tissue at one time point in the feeding cycle, still did not have sufficient coverage to adequately represent a complete record of salivary gland transcription.

To address these hypotheses I therefore undertook more in-depth comprehensive analysis using additional software, and generated and analyzed additional cDNA sequences from the same RNA samples but without size selection.

© University of Pretoria

Chapter 4

Given that tick genomes are large and more akin in size and organization to vertebrates than the model arthropod *Drosophila melanogaster* according to re-association kinetic data and sample sequence data (in the range of 2-7 gigabases as compared to 200 Mb) I hypothesized that this analogy might also extend to the transcriptome. In particular it has become apparent that many non-coding sequences are transcribed in vertebrate genomes (Mattick, 2003).

To test this postulated functional analogy between tick and vertebrate genomes, I therefore searched RaGI for putative non-coding transcripts and found evidence to support pseudogenes identified in the dataset.

Chapter 5

Ticks transmit the widest variety of disease-causing organisms amongst arthropods yet functional roles of tick proteins are not as well understood as those of other arthropods including mosquitoes and *Drosophila*. More than 60% of sequences in RaGI were not assigned a function in the initial annotation due to lack of significant sequence similarity with known proteins.

I compare transcribed sequences of four ixodid ticks so as to identify previously unknown functional proteins that are conserved between ticks and are likely to play important roles in tick biology.

Chapter 6

Initial investigations revealed the existence of a tRNA SINE-like family of transposable elements in *R. appendiculatus* (Sunter et al., 2008). I therefore hypothesized that the very large size of Ixodid tick genomes, including that of *R. appendiculatus*, could be partially attributable to the presence of additional families of transposable elements.

To test this hypothesis I made a focused search of the *R. appendiculatus* genomic and transcribed sequence data for different categories of transposable elements to describe the composition of the repetitive elements present in the genome, highlighting the abundance of Class I retrotransposons.

© University of Pretoria

Chapter 1. Literature Review

1.1 Tick Biology and medical and economic impact

In the tropics ticks are the second most important arthropod vectors for human pathogens - the first being mosquitoes. Ticks transmit the widest variety of pathogenic microorganisms amongst arthropods. Tick-borne diseases are a major constraint to livestock agriculture. A recent estimate showed the annual cost of tick-borne diseases to farmers in eastern, central and southern Africa to be US$168 million (Minjauw and Mcleod, 2003). Ticks of the genus *Ixodes* transmit numerous diseases to both humans and animals, including Lyme disease (caused by the spirochaete *Borrelia burdorferi)*, babesiosis (caused by the apicomplexan protozoan species in the genus *Babesia*) tick-borne encephalitis (caused by tick-borne encephalitis virus – TBEV) and human granulocytic anaplasmosis (caused by the rickettsial bacteria *Anaplasma phagocytophilum*).

1.1.1 Hard Ticks

Ticks are obligate haematophagus ectoparasites that are important vectors of a number of animal diseases, some of which are zoonotic. They parasitize vertebrates including mammals, reptiles, amphibians and birds. There are 889 currently known species of ticks in 22 genera. They are classified into three families. The ixodidae (hard ticks) that have thick exoskeletons made of chitin, the Argasidae (soft ticks) that have a membranous outer surface and Nuttalliellidae, which contains one rare African species, *Nuttalliella namaqua* found only in South Africa and Tanzania.

80% of the world's ticks belong to the *Ixodidae*, which contains 713 species (Barker and Murrell, 2004). Within the *Ixodidae* 249 species belong to the *Ixodes* genus (Barker and Murrell, 2004). The most economically important members of the *Ixodidae* are the genera *Amblyomma*, *Dermacentor*, *Haemaphysalis*, *Hyalomma*, *Ixodes*, *Rhipicephalus* and

© University of Pretoria

Boophilus, which are vectors of livestock and human diseases, certain species transmitting both livestock and human infective pathogens. In the case of *Rhipicephalus appendiculatus* it transmits both *Theileria parva*, the cause of East coast fever in cattle, and *Rickettsia conorii* that causes tick typhus in humans.

The majority of the tick species parasitize wild vertebrates, except for approximately 50, which parasitize humans, livestock and other domestic animals. Nearly all tick species spend most of their lives off the host. The parasitic phases of the life cycle are short compared to their long non-feeding (also known as fasting) periods. Ticks maintain their water balance efficiently in both these phases. Maintenance of water balance is fundamental to tick distribution, survival and pathogen-transmission (Anderson, 2002). When feeding, the tick anchors to the host's skin using its hypostome, which is attached by cement secreted in tick saliva, to create a feeding lesion that forms around its mouthpart. The cement forms a cone, which seals off the feeding lesion and protects the hypostome during feeding. Feeding causes direct physical damage to the host's skin and also induces inflammatory responses. Hard ticks feed over several days; most larvae feed for 3-5 days, nymphal instars 3-8 days and adults 6-12 days (Anderson, 2002). In the initial phase of the feeding tick reproductive tissues, salivary glands and the cuticle undergo development to enable expansion during the engorgement stage. A fully engorged female of a hard tick species may increase her body weight by upto 125 to 150 times (Anderson, 2002).

The feeding process involves an exchange of fluids between the tick and the host. The tick ingests blood, lysed tissues including components of lymphatic from the host and releases saliva, faeces and semi-digested blood. The saliva is injected into the host's tissue whereas the faeces and semi-digested blood are deposited on the skin of the host. The disease-causing pathogens are ingested or expelled during this phase. The midgut, where blood is digested, presents conditions that are favourable for survival of the pathogen in the tick compared to blood-feeding insects, in that the digestive proteases in ticks are intracellular having limited effect on ingested pathogens (Anderson, 2002).

© University of Pretoria

This characteristic in ticks also makes them particularly suitable as vectors of virus infections (Labuda and Nuttall, 2004).

To minimise the impact of host rejection responses and to prolong successful feeding tick saliva possesses properties that suppress immunological responses, haemostasis and inflammatory pathways in the host. Nevertheless, some mammalian hosts have been observed to develop resistance to repeated infestations by ticks. The mechanism of this resistance is presumably based at least partially on acquired immunity that neutralises tick molecules that modulate the host pathways mentioned above.

Tick life cycle comprises four stages – egg, larva, nymph and adult. They feed three times in their life – in the larval, nymphal and adult stages. They can have one, two or three - host life cycles, determined by the number of different hosts infested by the larval, nymphal and adult instars. One-host ticks attach to the vertebrate host at the larval stage and moult twice on the same host animal generating both nymphal and subsequently adult stages. Two-host ticks moult from larva to nymph on one host, disengage from the first host after blood feeding and moult into the adult instar free in the environment and subsequently locate a second host and attach. The second host may be of either the same or a different species than the first host. Three-host ticks undergo no moulting on the host animals. The larva attaches to the first host; it drops off to moult into a nymph, which attaches to a second host to feed. After engorgement the nymph disengages and moults into the adult instar which subsequently finds a third host to engorge on, prior to egg laying.

The generation time in ticks is often 1 to 2 years, but in extreme climates ticks can live as long as 6 years. Mating in adult ixodid ticks takes place on the host. After engorgement, the female drops off the host, lays eggs and then dies. The male remains on the host after mating for longer periods amounting to months. Egg batches of one-host ticks are typically smaller in number compared to those of two-host ticks and three-host ticks, which is likely correlated with the fact that one-host ticks have a higher probability of survival than the two tick types with more complex life history strategies.

© University of Pretoria

The engorgement weight of the female tick correlates with the number of eggs she lays (Wang et al., 2001).

Some ticks have a narrow range of host preference, while others feed on many different species. Most tick species have a preference for wild vertebrates. A relatively small percentage have successfully adapted to domestic animals and humans as hosts, resulting in many species transmitting disease-causing pathogens.

Ticks locate their host using either active or passive strategies. Passive tick species remain in their habitat in wait for their vertebrate host to invade. Active strategies to detect presence of the host include use of signals from the host, such as carbon di-oxide and ammonia over long distance, and vibration, body heat, and odourants such as lactic and butyric acids in hosts at close range. Some tick species also respond to sounds generated by hosts, for example, *R. sanguineus*, the brown dog tick is attracted to the bark of dogs (Waladde and Rice, 1982). In addition to host stimuli, pheromones released by feeding ticks on infested animals also act as attractants (Norval et al., 1989). On sensing these signals the tick crawls or runs in a suitable position for attaching to the host.

1.1.2 Economic Importance of Ticks

Livestock are a direct source of meat, milk and hides. They also provide traction and manure, which are key to the sustainability of mixed crop and animal agricultural systems, particularly in the developing world. They are particularly important for food security in poor countries, whose economies remain dominated by agriculture. Ticks and have been estimated to be responsible for US$ 17 billion of economic loss in the livestock (de Castro et al., 1997).

Ticks transmit more pathogens than any other haematophagous arthropods. They are vectors for important livestock diseases including theileriosis, babesiosis, anaplasmosis and cowdriosis (heartwater). Tick-borne diseases induce a spectrum of outcomes, including low productivity, reduced fecundity and reduction in live weight, decreased

© University of Pretoria

value of hides particularly as a result of infestation with species in the genus *Ambylomma*, and frequently death in livestock. Rapid mortality is a frequent outcome of severe East Coast Fever induced by *T. parva* infection in the ear of exotic *Bos taurus* cattle breeds, with an estimated one million deaths per year in East and Central Africa. Therefore, improved methods to control tick borne disease, which is currently effected by the increasing unsustainable application of chemical acaricides (Bishop et al., 2009) represents an important research goal. The most important disease-causing tick species in East Africa are *Rhipicephalus appendiculatus,* the major vector of *T. parva* (Norval et al., 1992), *R. decoloratus* and *R. microplus* that are vectors of *Anaplasma and Babesia*, *Amblyomma* spp. that are vectors of the rickettsia *Ehrlichia ruminantium,* and *Hyalomma* spp. that transmit zoonotic viruses and bacteria to humans.

Ticks co-exist with wild animals in nature without apparent serious adverse health impacts on their hosts; however, they become problematic for domestic livestock when either tick-naive livestock are moved into areas where ticks are endemic, which can be frequent when attempts are made to increase livestock productivity, or when tick-infested animals are introduced into regions where ticks and their associated pathogens were not previously present.

1.2 *Rhipicephalus appendiculatus* and its role in causing East Coast Fever in cattle

Rhipicephalus is one of the largest genera of hard ticks. It includes some of Africa's most economically significant species. There are 79 species in this genus, including the five *Rhipicephalus* (*Boophilus*) species that have recently been integrated in a single genus as a result of analysis of molecular data (Murrell and Barker, 2003). Most of these species are found on a wide range of mammalian hosts. Cattle are the major hosts, amongst domestic animals, for the nymphal and adult stages of *Rhipicephalus*, whereas larval instars frequently feed on smaller mammals such as lagomorphs. *R. appendiculatus* is the most important of the *Rhipicephalus* species in East and Central Africa. It is also

© University of Pretoria

known as the African brown-ear tick as it prefers to attach to and feed on the ears of its animal host in the adult stage. Heavy infestation of *R. appendiculatus* on the ears of the host severely damages the ears, and can sometimes result in fatal toxaemia.

Rhipicephalus is paraphyletic with respect to *Rhipicephalus* (*Boophilus*) (Murrell and Barker, 2003), a phylogenetic arrangement which was confirmed by comparing 30 morphological characteristics as well as sequences derived from 12S rRNA, 18S rRNA, cytochrome oxidase I and internal transcribes spacer 2 genes of 33 species from the Rhipicephalinae and Hyalomminae tick subfamilies (Murrell et al., 2001). The same study provided evidence to suggest that subfamily Rhipiciphelinae is also paraphyletic with respect to Hyalomminae. However, this change in nomenclature has not yet been adopted by the tick research community.

R. appendiculatus is a three-host tick that parasitizes mainly cattle amongst domestic animals and several wild ungulate species including buffalo, giraffe and waterbuck (Norval et al., 1992:109), but is capable of feeding on more than a hundred vertebrates. Resistance to *R. appendiculatus* has been observed in both domestic and wild animals. *R. appendiculatus*-resistant domestic cattle include Zebu (originating from South Asian *Bos indicus*) and the southern African Sanga breeds (which contain a mixture of alleles from both the *Bos taurus* and *Bos indicus* lineages), which are less susceptible than exotic *Bos taurus* derived breeds and their direct crosses with taurine cattle that were originally domesticated in the Middle East. Amongst wild animals, wildebeest and warthog appear to be highly resistant to the tick (Lightfoot and Norval, 1981).

The distribution of *R. appendiculatus* depends on several factors. Most important of which are climatic conditions, vegetation and host availability. Cool, shady savannah areas with a minimum of 24 inches of annual rainfall, good vegetation cover and populations of suitable host animals are necessary conditions for the tick to survive.

R. appendiculatus is the primary host for *Theileria parva*, an intracellular apicomplexan protozoan parasite that causes East Coast Fever (ECF). ECF is a fatal lympho-

© University of Pretoria

proliferative disease of cattle, occurring in the sub-Saharan Eastern, Central and Southern Africa and resulting in death in susceptible taurine animals between three to four weeks post-infection. The disease has been present in cattle in the afflicted region of Africa for several hundred years, although the exact timing of transmission from cape buffalo (*Syncerus caffer),* the major wildlife reservoir, to cattle, is not known. The association of *T. parva* with *R. appendiculatus* is probably much more ancient given the greater diversity of the parasite in buffalo as compared to cattle (reviewed by Bishop et al., 2009). More than 25 million cattle are currently threatened by *T. parva* in countries within the regions of Africa where infected ticks occur.

East Coast fever is a major constraint limiting improved livestock production in East Africa causing >80% mortality in adult exotic cattle. *R. appendiculatus* is also the principal vector of Nairobi sheep disease virus, a sheep and human pathogen, and Thogoto virus, which has also been associated with disease in sheep and humans. In addition to transmitting other species of *Theileria,* particularly the non-pathogenic *T. taurotragi* which originates from Eland (*Taurotragus oryx*), and also the rickettsia, *Ehrlichia bovis* to cattle and other large ungulates, *R. appendiculatus* is a vector of *Rickettsia conori*, causing tick-bite fever in man. As well as acting as a vector of pathogens, high infestations with adult *R. appendiculatus* (which can exceed 1000 on cattle) may give rise to toxicosis in the host resulting in immune-suppression and the recurrence of tick-borne diseases to which the animals were previously immune. Heavy infestations in calves can also cause death resulting from acute and chronic anaemia. *R. appendiculatus* is an excellent model for tick research as it can be readily maintained on laboratory animals and has been used extensively in studies of the promotion of tick-borne virus transmission by molecules secreted from *R. appendiculatus* salivary glands (Nuttall, 1998). *R. appendiculatus* has also been employed to investigate differences in vector competence for *T. parva* infection between tick stocks and the heritability of this phenotype (Ochanda *et al.*, 1998; Young *et al.*, 1995).

© University of Pretoria

1.2.1 The life cycle of *T. parva*

To briefly summarise the life cycle, when an adult tick bites an infected animal, it ingests *Theileria* piroplasm-infected red blood cells together with the blood meal. Most piroplasms are destroyed in the midgut however a few survive allowing gametogenesis to occur, which results in two structurally similar haploid gametes, which fuse to form a diploid zygote (Gauer et al., 1995). The zygote then invades a gut epithelial cell and differentiates into a motile uni-nucleate kinete in the gut epithelium. Kinete differentiation occurs in synchrony with tick moulting (Young and Leitch, 1980). After the ticks have moulted to the next instar the kinetes invade the type III acinar cells in the tick salivary glands where sporogony, which is maturation of kinetes to the host-infective sporozoite stage, occurs. This process results in a 25 fold increase in the mass and protein content of the female salivary glands (Bowman and Sauer, 2005). From days 3-7 during subsequent feeding sporozoites are gradually released into the tissues of the mammalian host.

Theileria sporozoites infect host B cells, T cells, other lymphocytes or monocytes, depending on the species. In case of *T. parva*, infections are primarily in lymphocytes (Baldwin et al., 1988). Following a rapid energy-dependent, receptor-mediated process of entry into the host lymphocyte (Shaw, 1997), sporozoites emerge from the parasitophorous vacuole within 30 minutes and exist free in the host cytoplasm, unlike *Plasmodium* species but similar to several phylogenetically unrelated pathogens including the protozoan *Trypanosoma cruzi* and the bacterium *Listeria* (Andrews and Webster, 1991). They then differentiate into multinucleated schizonts. The schizonts induce a state of permanent activation of the infected lymphocytes, which exhibit a phenotype similar to that of tumour cells. As the lymphocyte undergoes cell division, the schizont divides in synchrony, by associating with the host mitotic spindle, resulting in two infected daughter cells. This clonal expansion results in an approximately ten-fold increase in schizont-infected cells every three days. After approximately 14 days, schizonts undergo intra-lymphocytic merogony within the lymphocytes, to generate merozoites. Rupture of the host cells releases merozoites, which invade blood cells and

© University of Pretoria

differentiate into unicellular piroplasms. Uptake of piroplasms during tick feeding and subsequent development within the tick gut completes the *T. parva* life cycle.

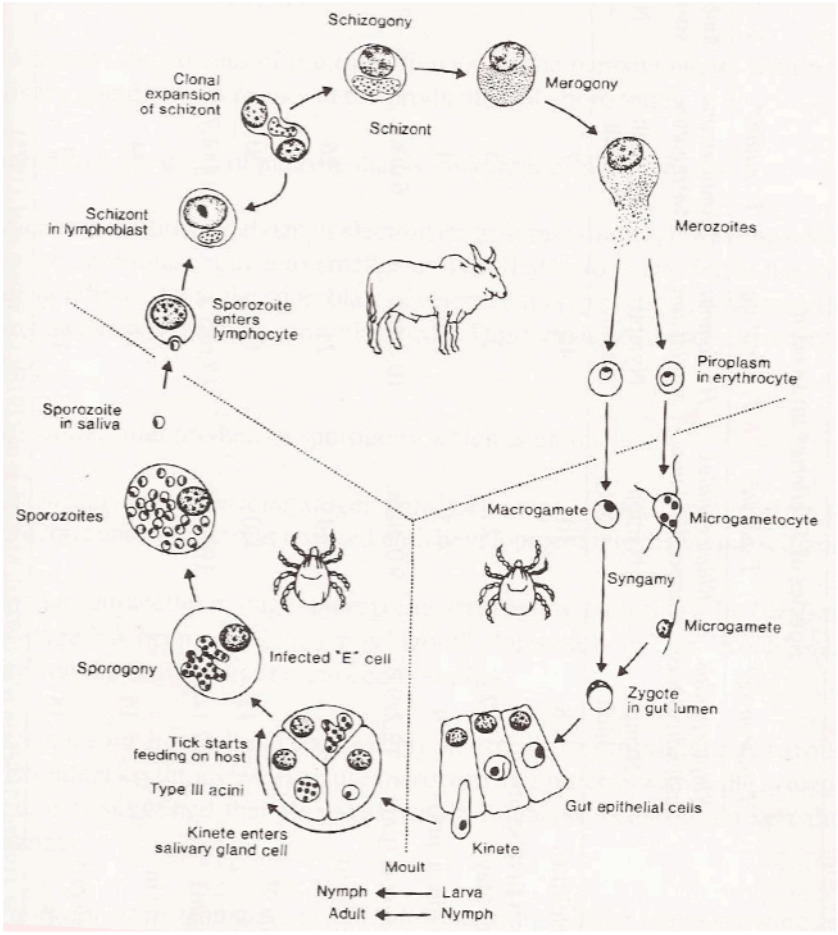


Figure 1.1 Life cycle of Theileria parva in the tick vector and animal host. (Picture source: Norval et al., 1992)

R. appendiculatus can be infected by, and transmit, genetically different *Theileria* taxa. *T. parva bovis* and *T. parva lawrencei* originating from cape buffalo resulting in January disease and Corridor disease, respectively, in cattle. The form of the disease originating from the buffalo reservoir has a different clinical manifestation as compared to East Coast Fever, resulting in rapid death with low piroplasm parasitaemia and schizont parasitosis. Additionally, *T. parva* can coexist in the salivary glands of *R. appendiculatus* with *T. taurotragi*. As mentioned above, *Rhipicephalus* species are vectors for several viral diseases in livestock and humans (Labuda and Nuttall, 2004). *R. appendiculatus* is a vector for Nairobi Sheep Disease virus, Dugbe virus and Thogoto virus.

© University of Pretoria

1.3 Tick control

1.3.1 Acaricide

For centuries, tick control has been effected using chemical acaricides. Dipping in arsenical compounds was used against *R. microplus* and *R. annulatus* and worked well to eradicate these tick species from the USA in early 1900s. But resistance subsequently developed in the ticks. Organochlorine insecticides such as DDT and Benezenehexachloride, introduced later, eliminate ticks by preventing acetylcholine binding to its receptor, hence over-stimulating the sodium channels in neurons. However, resistance to organochlorine insecticides also developed in *R. appendiculatus, R. microplus and R. decoloratus* in Australia and Africa. Furthermore, the residues persisted in the environment with potential but as yet unquantified implications for wildlife populations and human health. As a result of these factors their use has been discontinued. Organochlorines were replaced by organophosphates and organocarbamates, mainly to control *Rhipicephalus* (*Boophilus)* ticks. They function by inhibiting acetylcholinesterase thereby inducing continuous nerve firing. Unlike organochlorines, they do not persist in the environment, however, their toxicity to vertebrates, combined with emerging resistance in ticks has led to a decline in use. Amitraz, a member of the formamidine chemical family, is used to control a wide range of invertebrates and organophosphate-resistant ticks, including *R. microplus, R. decoloratus, R. appendiculatus and R. evertsi* on cattle and other domestic animals. Ticks on treated animals are usually killed either prior to attachment or within 24 hours of attachment. Pyrethroids are synthetic compounds that, like most insecticides, affect the nervous system of the invertebrate. They are costly but effective. Benzoyle phenyl ureas such as Fluazuron, inhibit chitin formation in *B. microplus*, which in turn leads to a decline in the fecundity and fertility of engorged female ticks. Due to its lipophilic property Fluazuron is excreted in milk, transmitting chemical protection to the calves. However, the meat of such cattle cannot be consumed until the residues of the chemical have waned from the animal's fat tissues. Spinosad confers about 90% control of *R. microplus*. It functions by binding to the nicotinic acetylcholine receptors on the

© University of Pretoria

postsynaptic cell membrane, and is effective against all developmental stages of the tick. This acaricide is a fermentation metabolite of an actinomycete fungus.

Regimes used for acaricide administration are determined by several factors, including chemical stability, mode of action, toxicity, pharmacokinetics in animals and cost. Methods of application of acaricidal control include dipping vats, spray races, hand spraying, pour-ons, direct injection, intra-ruminal boluses, acaricide-impregnated ear-tags and pheromone/acaricide-impregnated devices.

For acaricide treatment to be effective it has to be administered consistently in order to ensure sustainability. This is costly and time consuming. The cost of acaricide application to cattle in East Africa is estimated to be between US$ 6-36 per cattle per year (Minjauw and Mcleod, 2003). Acaricide application not only encourages development of resistance in ticks, it also suppresses acquisition of natural immunity to the ectoparasitic tick and also to tick-transmitted diseases.

Since resistance to acaricides is increasing (Graf et al., 2004) and evolving more rapidly with each new product that is released, and additionally, there are concerns regarding residual toxins in milk and meat, alternate strategies of tick control are being explored to reduce the spread of resistance. One of these strategies involves sequential application of one chemically distinct acaricide targeting different host pathways on a rotational basis to discourage evolution of acaricide resistance in the ticks. An additional strategy for tick control practiced is the use of tick-resistant cattle. However, there is likely to be a trade off with production traits, such as milk yield, which has not yet been fully quantified.

1.3.2 The Infection and Treatment (ITM) live Immunisation Strategy

Cattle, particularly indigenous *Bos indicus* breeds that have survived infection with *Theileria*, acquire long-term immunity to clinical theileriosis, although not necessarily against super-infection with *T. parva* genotypes that differ from the immunizing strain.

© University of Pretoria

This implies that immunizing cattle could be an effective measure of protection against the disease. Initial unsuccessful attempts to immunize against ECF in the early 1900s using blood and material from spleen and lymph nodes of infected or recovered cows resulted in fatal ECF in the animals. However, by concurrent administration of the antibiotic tetracycline, to suppress clinical symptoms during the incubation period of ECF, it was observed that mortality was reduced and animals were successfully immunized (Neitz, 1953). The method was subsequently refined through the use of sporozoite stabilates derived from cryopreserved, homogenised whole *T. parva*-infected ticks and a suitable long acting formulation of tetracycline (Radley et al., 1975a). This was known as the Infection and Treatment Method (ITM).

Success in the ITM was later further improved by standardising production of a supernatant of homogenised infected adult ticks along with an increased dose of tetracycline to limit the numbers of severe reactors. In addition, in order to widen the protection produced by the ITM procedure the use of a 'cocktail' of *T. parva* stocks was adopted. These measures were shown to confer broad protection (Radley et al., 1975b). A theoretical concern in using a standard 'cocktail' is that if used for immunizing outside the region of origin of the strains making up the cocktail, there lies a risk of introducing novel antigenic types to the environment, since immunized animals become persistently infected 'carriers', that have been demonstrated to transmit infection to local ticks and cattle. However, due to the complex mix of *T. parva* genotypes already present this does not pose a serious risk in endemic areas but only in disease-free zones that are adjacent to the endemic regions. If the *R. appendiculatus* is present in the former, the risk of spreading the disease to new areas increases. Immunization by ITM must be performed prophylactically and cannot be used as a therapeutic strategy. Additional challenges to this method are presented by the need for a cold chain, which limits deployment in areas with poor infrastructure. Despite these limitations the method has been successfully deployed in certain production systems, particularly in Maasai pastoralist systems in Tanzania and Southern Kenya. The current status has been reviewed by Uilenberg (1999) and Di Giulio et al. (2009) .

© University of Pretoria

1.3.3 Approaches to Development of Subunit Vaccines for Tick Vectors and the Parasites That They Transmit

Subunit vaccines can confer immunological protection through rapid induction of memory responses against tick infestations and parasite challenge in immunised animals. They are environmentally friendly, more cost-effective and are theoretically less likely to evolve resistance as compared to chemical acaricides. They can also potentially target the tick vector directly or enhance the performance of an anti-pathogen vaccine. The first ever vaccine against a parasite was developed to control the tick *Rhipicephalus* (then *Boophilus*) *microplus* by Willadsen and co-workers (Willadsen et al., 1995) using a hidden gut antigen designated BM86. Effects of the vaccine can be measured in the vector or in the host. In the tick vector, vaccine effects include decrease in engorgement weight, number of eggs laid, duration of attachment and feeding and mortality. The host immune response is used as a measure of vaccine efficacy in the host. Experimental vaccines based on BM86 and its homologues have been demonstrated to be effective against a range of related tick vector species.

Vaccines can potentially target multiple antigens that are both exposed and concealed. Exposed antigens are those that induce a natural host immune response on tick infestation whereas concealed antigens are not part of the natural tick-host interaction and therefore do not induce an immune response under normal circumstances. Although the latter require regular boosting by re-vaccination to maintain a response to challenge in the host, with BM86 representing a good example of this, a corollary of this is immune evasion of the vaccine, which is theoretically less likely to evolve in the host.

There are several strategies for identification of antigens for incorporation in vaccines to control tick infestation. Antigens have been identified by studying the antibody responses of an immune host to tick antigens. The p29 antigen from *Haemaphysalis longicornis* was discovered in this manner (Mulenga et al., 1999). Immature and mature *H. longicornis* ticks were fed on rabbits until resistance to infestation was observed. Serum from the rabbits was collected and used to immuno-screen a cDNA library

© University of Pretoria

constructed from RNA extracted from the tick salivary glands. Molecular analyses led to identification of the p29 antigen. p29 is a 29 kDa salivary gland-associated protein, 277 amino acids long containing a putative signal peptide sequence. The antigen exhibits sequence similarity with numerous glycine-rich extracellular matrix proteins. It has structural similarities with collagen-like proteins. Vaccination with recombinant p29 (rp29) resulted in up to 40% mortality in larval ticks and 56% in nymphal ticks using a homologous challenge (Mulenga et al., 1999). A reduction in engorgement weight was also observed in adult ticks.

Antigen recognition in proteins hypothesised to be of functional importance for tick survival has also led to identifying vaccine candidates. Sugino and colleagues (Sugino et al., 2003) isolated serpin HLS1 from *H. longicornis* by priming synthesis of tick cDNAs from total RNA of ticks, with nucleotides designed from conserved amino acid domains located within serpins. Approximately 44% nymphal and 11% adult tick mortality was recorded on challenge after vaccination of hosts with recombinant HLS1 (rHLS1). This method of antigen identification is limited by the fact that knowledge of functionally important tick proteins is incomplete and that not all tick proteins identified will induce immune responses when used to vaccinate the host.

The third, and the most direct method for antigen identification, is the use of biochemical fractionation of protein mixtures and subsequent use of these to experimentally vaccinate host animals in combination with adjuvant followed by tick challenge. This is a lengthy and labour intensive but systematic and effective procedure. As mentioned previously this method was used successfully to identify the concealed antigen Bm86 from *R. microplus* (Willadsen et al., 1988). Antibodies to Bm86 inhibit the endocytotic activity of digestive cells in the tick gut that bring about endocytosis and digestion of the blood meal (Willadsen et al., 1989). The Bm86 protein is 650 amino acids in length and has a predicted molecular weight of 71.7 kDa. The extensive glycosylation of the protein probably contributes to strong antibody response induced by the molecule following biochemical purification from the tick gut tissues. However, vaccination studies using glycosylated baculovirus recombinant and non-glycosylated

© University of Pretoria

bacterial recombinant forms of the protein indicate that the antibody responses are not essential for the efficacy of the subunit vaccine, although antibodies are believed to be the main element of the protection induced (Willadsen et al., 1995). Bm86 has been found to be expressed in all stages of the tick's life cycle (Willadsen, 2004). Cross-protection of Bm86 was best exhibited against *R. annulatus* challenge (de Vos et al., 2001), whereas the *R. microplus* BM86 vaccine exhibited no protection against *R. appendiculatus* and *A. variegatum* challenge. The Bm86-based vaccine manufactured in Australia has been marketed since 1994 under the trade names TickGARD and TickGARD Plus. Another vaccine based on the same antigen isolated from a different tick isolate was manufactured in Cuba and sold under the trade names Gavac and Gavac plus. Gavac has been used extensively in Latin America and demonstrated to reduce acaricide usage when deployed as a component of integrated control strategies, particularly in Mexico (de la Fuente et al., 1998). Both TickGARD and Gavac control tick numbers over successive generations, primarily by reducing fecundity. However both currently require booster vaccinations on a six monthly basis. In one sense they can therefore be considered environmentally benign acaricides.

1.4 Tick salivary gland function and modulation of host pathways

Tick salivary glands play a major role in tick feeding (McSwain et al., 1982). They are responsible for maintaining homeostasis in the tick, provide lubrication for the mouthparts, in many ixodid species they producing tick cement which assists in holding the mouthpart in place while feeding. They produce immuno-modulatory molecules that modulate the host haemostatic, inflammatory and immune responses (Francischetti et al., 2010; Ribeiro and Francischetti, 2003). This facilitates tick feeding by minimising tick rejection, which is frequently observed when a tick feeds on a non-natural host (Ribeiro, 1989). As host immune responses to tick feeding can protect against pathogen transmission (Valenzuela, 2004), so the suppression of host pathways by the tick can also promote pathogen transmission.

© University of Pretoria

Because of the important role salivary glands play in modifying host immune responses, salivary gland proteins have recently been the subject of in-depth investigation. Complementary DNA (cDNA) libraries have been constructed from the salivary gland tissues of several tick species including the genera *Amblyomma* (Aljamali et al., 2009; Nene et al., 2002), *Dermacentor* (Alarcon-Chaidez et al., 2007), *Ixodes* (Chmelař et al., 2008; Ribeiro et al., 2006; Valenzuela et al., 2002a), *Rhipicephalus* (Anatriello et al., 2010; Lambson et al., 2005; Nene et al., 2004; Santos et al., 2004), *Haemaphysalis* (Nakajima et al., 2005) and *Hyalomma* (Francischetti et al., 2011) among the ixodid hard ticks and the genera O*rnithodoros* and *Argas* among the argasid soft ticks (Francischetti et al., 2008a; Francischetti et al., 2008b; Mans et al., 2008b). Analysis of salivary gland transcriptomes revealed that the most abundantly expressed proteins belonged to multi-gene families, for example, five homologues of the Collagen-like Secreted Proteins (CLSP) AAT92166.1 (*I. pacificus*) and four of AAM93621.1 (*I. scapularis*) were present in *I. ricinus* (Chmelar et al., 2008); and in *A. variegatum,* glycine-rich proteins, shown to be important for tick attachment (Kemp et al., 1982), made up the most abundant sequences, and of which 11 families were identified (Nene et al., 2002). Differential expression of some of these proteins was also observed. For example, expression of histamine binding proteins (HBP) in female *I. ricinus* increased with the days fed (Chmelař et al., 2008); basic tail proteins (thought to have anticlotting function) were overexpressed in *I. scapularis* nymphs compared to adults and in adult females 18-24 hours post attachment (Ribeiro et al., 2006). A large proportion of the transcripts sequenced had no known database matches or putative functions, for example, 35% of *I. ricinus* (Chmelař et al., 2008), 61% of *A. variegatum* (Nene et al., 2002), 60% of *R. appendiculatus* (Nene et al., 2004) and 15% of *I. scapularis* (Ribeiro et al., 2006) transcripts from the salivary glands had no matches to known proteins.

More than 18,000 salivary gland transcripts were sequenced from 4-day fed *R. appendiculatus* ticks uninfected and infected with *T. parva* (Nene et al., 2004). The transcripts consisted of proteins from *R. appendiculatus*, or their variants, that have been shown to be of importance in tick biology or confer partial protection against the

© University of Pretoria

tick in immunized animals, such as HBPs, Serpins, Trp64 and RIM36. Proteins homologous to other ixodid tick species were also present. No significant differential expression between the infected and uninfected libraries was observed.

The work described in this thesis builds on the initial analysis of Nene et al. (2004) by providing an in-depth analysis of the salivary gland transcriptome of *R. appendiculatus*.

© University of Pretoria

Chapter 2. Materials and Methods

2.1 Preparation of tick material

In brief, adult male and female ticks were fed on uninfected Boran cattle and allowed to mate. Engorged females were collected and maintained under optimum conditions for laying eggs. Subsequent to hatching of the eggs, larvae were left to harden before feeding them on uninfected rabbits. The nymphs that moulted from engorged larvae were allowed to feed on a Boran calf that had been inoculated with the *T. parva* Muguga sporozoite stabilate 3087. The nymphs were attached to the calf as soon as piroplasm parasitaemia was detected. They were fed to engorgement and incubated to allow them to moult into adults. Adult ticks were then fed on rabbits for four days before their salivary glands were isolated by dissection to prepare a cDNA library of infected ticks labelled RAB. Similarly, a cDNA library was constructed from uninfected ticks and was labelled RAA.

2.2 EST library construction

2.2.1 by life Sciences (with size selection by gel purification)

As described in Nene et al., (2004), total RNA was extracted from salivary glands of approximately 500 ticks. Following RNA extraction and size selection of RNA > 2 kilobases in length by purification from an agarose gel, a directional cDNA library was constructed by priming with an oligo(dT) primer containing a *Not*I restriction site. The single stranded cDNA was converted into double stranded cDNA by reverse transcription, and cloned into the plasmid vector pCMVSPORT6.ccdb between *Not*I and *Eco*RV sites. The vector was transfected into *E. coli* cells and amplified. Plasmid template was prepared and sequenced. 5' ESTs were sequenced using a vector primer whereas 3' ESTs were primed using oligo-$dT_{23}V$ (where V=dA, dC or dG) primer.

© University of Pretoria

2.2.2 NIH library (without size selection of RNA)

The library was prepared at the National lnstitutes of Health (NIH) as described in Chmelar et al. (2008), from RNA isolated from salivary glands of uninfected ticks, described above. 20 x 96 clones were sequenced using the Sanger method, from which 1784 good quality sequences were obtained.

2.3 Clustering and Annotation

2.3.1 *R. appendiculatus* Gene Index (RaGI)

ESTs were then processed *in silico* to construct the gene index database. ESTs from sequences from both the *T. parva*-infected and un-infected salivary gland libraries were pooled, the vector, adaptor and contaminant sequences removed, and the polyA/T tail trimmed. *R. appendiculatus* gene sequences were downloaded from GenBank (2004). The mix of sequences was clustered using stringent criteria to group similar sequences. Sequences in a specific cluster were defined as being more than 95% identical over a region of more than 40 nucleotides and having less than a 20 bp mismatch at either end. Each cluster was assembled separately using CAP3 (Huang and Madan, 1999) to generate tentative consensus sequences (TCs). ESTs that couldn't be assembled were termed singletons.

Automated annotation of the assembled sequences and singletons was performed by searching the GenBank non-redundant protein database, referred to as NR hereon (which includes translated annotated coding regions from GenBank nucleotide database, RefSeq, SwissProt, PIR, PRF and PDB; version unknown) using an E-value cut-off of 1e-25 for the BlastX algorithm. This assigned a tentative function to sequences with hits above the threshold. It is worth noting that no manual curation of the data was performed at this stage.

© University of Pretoria

2.3.2 NIH library

The following were carried out by collaborators at the National Institutes of Health (NIH). EST sequences were trimmed of vector sequence and polyA/T tails. Sequences containing > 5% unassigned nucleotide ('N') were discarded. In total 1638 ESTs were assembled into 1202 contigs and singletons.

For assembly ESTs were clustered first using BlastN (minimum identity > 95% over 64 bp). The resulting clusters were assembled using CAP3. To determine if novel transcripts had been captured the assemblies were searched against the RaGI database using BlastN (-W 40). Matches with > 96% identity over > 100 bp or 94% identity over > 300 bp were not considered as 'novel'.

Protein homologues for the contigs were sought using BlastX (E-value < 1e-3; version 2.2.18) against NR (GenBank Release 168.0; RefSeq Release 31), Swissprot (Release 14.0), mitochondrial and rRNA sequences from GenBank. A specialized database of sequences from Ixodid species was constructed and searched for protein homologues.

2.4 BAC library preparation

2.4.1 Preparation of *R. appendiculatus* DNA

Frozen eggs of the *R. appendiculatus* Muguga stock that has been maintained at ILRI for over 40 years were ground and re-suspended in PBS (phosphate-buffered saline). After two washes in PBS, the material was re-suspended in PBS and mixed with an equal volume of 1% low melting point agarose. Plugs were prepared from the cell/agarose mixture using a DNA plug mold kit (Bio-Rad). The plugs were incubated in 10 mM Tris–HCl/0.5 M EDTA (TE buffer), containing 1% N-lauryl sarcosine and 0.2% proteinase K at 50 °C overnight. Following dialysis in TE, the plugs were stored in TE at 4°C.

© University of Pretoria

2.4.2 Construction, sequencing and assembly of bacterial artificial chromosome (BAC) clones

A BAC library was constructed in the pECBAC1 vector by Amplicon Express (http://www.genomex.com) from agarose plugs containing high molecular weight DNA from *R. appendiculatus*. Partially digested DNA with *BamH*I was ligated into the *Mbo*I site of pECBAC1. Ligations were transformed into DH10b *E. coli* cells, and individual colonies were picked robotically and arrayed into 384 well plates. The library clones have an average insert size of 115 Kb, and based on a presumed genome size of 1×10^9 bp the library represents 5.5 X coverage of the genome. Three clones were randomly selected from the BAC library for nucleotide sequence determination. DNA sequencing and subsequent assembly was performed at TIGR using Celera (Myers et al., 2000) and TIGR (Sutton et al., 1995) assemblers, as described in Desjardins et al., (2007).

2.5 ITM stabilate sequencing and assembly

A sample from the ITM vaccine stabilate ILRI 08004 from the batch of 2008 was sequenced using the Roche 454 GS FLX Titanium Chemistry (manufacturers' protocol followed). The reads were assembled using Roche 454 *de novo* assembly software, Newbler GSAssembler (Release 2.0.00.20). Minimum overlap identity parameter (mi) settings tested were 90% (default) and 95%. The option for complex genome assembly was selected as the stabilate was expected to contain DNA from the tick vector, the mammalian host and the protozoan parasite, *Theileria parva*. For the final assembly on which downstream analysis was based, mi of 95% was used so as to reduce clustering of the different strains of *T. parva* that compose the ITM stabilate, although the default parameter settings would have been sufficiently stringent to separate sequences originating from the distant organisms present in the sample.

© University of Pretoria

2.6 Quantification of Ruka copy number in *R. appendiculatus* genomic DNA using quantitative real time PCR (RT-PCR)

2.6.1 Extraction of *R. appendiculatus* gDNA

Genomic DNA was extracted from *R. appendiculatus* eggs using the phenol-chloroform method described by (Sambrook et al., 1989).

2.6.2 Construction of calibration curves for Ruka sequences

An *E. coli* plasmid clone identified within the *R. appendiculatus* gene index (Nene et al., 2004), RAAAQ49TF, containing a single copy of Ruka, was obtained from the *R. appendiculatus* salivary gland EST library (Nene et al., 2004). This was used as a standard for the RT-PCR reaction. Plasmid was prepared using the Wizard Plus SV Minipreps DNA Purification System (Promega) according to the manufacturer's instructions. The purified plasmid DNA was linearised with the restriction enzyme *Not*I. Linearisation was confirmed by agarose gel electrophoresis on a 1.2% Seakem preparative gel. The linearised plasmid DNA was gel purified using a QIAquick Purification Kit (Qiagen) according to the manufacturer's instructions. Plasmid DNA concentrations were measured using a Nanodrop spectrophotometer (NanoDrop, Wilmington, DE) and the copy numbers were calculated using the following equation:

$$\text{Copies}/\mu\text{l} = \frac{(6.02 \times 10^{23}\ \text{copies}) \times (\text{plasmid concentration g}/\mu\text{l})}{(\text{Number of bases}) \times (660\ \text{Da/base})}$$

Five-fold serial dilutions of the plasmid (107 copies to 103 copies) in nuclease-free water were prepared. Single use aliquots of the standard dilutions were stored at – 80°C to ensure plasmid DNA stability (Applied Biosystems; Godornes et al., 2007). Genomic

© University of Pretoria

DNA standards were made by serial dilution of *R. appendiculatus* egg genomic DNA (2.3–23000 pg).

2.6.3 Quantitative Real Time PCR

Real time PCR was performed in triplicate with SYBR® GREEN PCR Master Mix (Applied Biosystems) with 300– 500 nM primers, in a final volume of 25 µl. PCR was performed using a 7500 Real Time PCR System (Applied Biosystems) for 40 cycles at 95°C for 10s, 60°C for 1 min, and 72°C for 2 min. Primer sequences that were conserved among several different Ruka elements were derived from an alignment of the eight most conserved Ruka sequences present in the sequenced RA BAC clones (Appendix Figure A2.3). The primer sequences are listed in Table 2.1. They were arranged into six pairs FWD_1 and REV_1 (Ruka 1), FWD_2 and REV_1 (Ruka 2), FWD_3 and REV_1 (Ruka 3), FWD_1 and REV_2 (Ruka 4), FWD_2 and REV_2 (Ruka 5) and FWD_3 and REV_2 (Ruka 6). A melting curve analysis was performed after the amplification phase, to check for non-specific amplification or primer–dimer formation.

The threshold cycle (Ct), the cycle number at which the fluorescence of the sample exceeded that of the background, was determined by 7500 Real Time PCR System sequence detection system version 1.2.1 (Applied Biosystems) using the standard curve method. Data analysis was performed using the same software.

Table 2.1 Primers used for quantitative real time PCR of Ruka from genomic DNA

Primer name	Sequence
FWD_1	5′-GYGGTTABGGBGCTCGRCTGCTGACC-3′
FWD_2	5′-GGBGCTCGRCTGCTGACCSGMAGGT-3′
FWD_3	5′-TCGRCTGCTGACCSGMAGGTHGCG-3′
REV_1	5′-TTATGAGRGACGCCGTAGTGGAGGGCT-3′
REV_2	5′-TGGGGTTTWACGTCCCAAAACCAC-3′

© University of Pretoria

2.7 Structure Prediction

2.7.1 Protein structure prediction

Six-frame amino acid translation of nucleotide TC sequences was done using the universal genetic code and Open Reading Frames (ORF) were predicted using Artemis (Rutherford et al., 2000). The longest ORF was assumed to be the coding part of the sequence. For secondary structure prediction of the amino acid sequence corresponding to the longest ORF a Position Specific Scoring Matrix (PSSM) algorithm implemented in PSIPRED v3.0 (Jones, 1999a) was used along with the Secondary Structure Element Alignment (SSEA) algorithm implemented in DomPred (Marsden et al., 2002). Fold prediction was carried out using the fully automated software GenTHREADER (Jones, 1999b).

I-Tasser (Zhang, 2008), which has been ranked as the best server for protein structure prediction in Critical Assessment of protein Structure Prediction (CASP) competitions 7 – 10 was used to derive tertiary structure models for ORFs having no closely related structures in the PDB. The model with the highest confidence score and a score > -5 was considered for further analysis. The model with TM-score > 0.5 was considered to have correct topology.

To see if the predicted models were structurally similar to any known protein structures the Dali server V. 3 (Holm and Rosenström, 2010) was used retaining hits with Z-score > 2.

2.7.2 Transfer RNA secondary structure prediction

tRNAscan-SE 1.21 (Lowe and Eddy, 1997) which was accessed via http://lowelab.ucsc.edu/tRNAscan-SE/ was used to identify a transfer RNA (tRNA) from a nucleotide sequence. The default search mode on eukaryotic source was used. It was also used for tRNA secondary structure prediction of Ruka.

© University of Pretoria

2.8 Identification of Glycine-rich proteins

First ORFs in all TCs and singletons of RaGI were identified using the Orfer software (Ribeiro J. unpublished), which takes an input of nucleotide sequence, and outputs a six-frame protein translation of the nucleotide. Amino acid composition of the protein translations was calculated in BioEdit. The output was parsed for conversion into a tab-delimited file using a Perl script.

Glycine-rich sequences (GRS) were identified if they had glycine content > 23%. This cut-off was based on the glycine content of 23.73% for RIM36, a well-studied GR protein. The protein translation was trimmed so as to retain a single ORF containing the glycine-rich region. Additional GRSs in RaGI were identified from keyword searches of assigned annotations. This was also done for GRPs in BmiGI, AvGI and IsGI to derive a list of GRPs in tick gene indices.

All GRSs identified were manually checked to eliminate short, low complexity sequences and those that were falsely annotated as Glycine-rich by automation. GRSs that were retained were checked to contain glycine-rich repeat patterns characteristic of cement proteins (Bishop et al., 2002).

2.9 Assessment of non-coding potential of a transcribed sequence using PORTRAIT

Homology-based ncRNA prediction software is not suitable for identifying ncRNAs in poorly characterized organisms. PORTRAIT (Arrial et al., 2009) is an algorithm for predicting ncRNA specifically in non-model organisms. It uses relaxed criteria for ORF prediction by tolerating sequencing errors and frameshifts in transcript sequence and using Support Vector Machines (SVM) it evaluates the coding potential of a transcript. Transcripts having scores greater than 50% were considered non-coding.

© University of Pretoria

The 7340 TC and singleton sequences in RaGI were subject to PORTRAIT analysis. TCs that were tentatively annotated as 'similar to', 'homologous to' and 'complete' matches to a known protein as well as had assigned GO annotations were used for benchmarking. 217 TCs fit the criteria. For 95% of the benchmarking TCs the non-coding probability was predicted to be > 50% with a percentage error rate of 5.34%. Less than 30% of the benchmarking TCs had a percentage coverage of coding ORF of > 75%, i. e. in > 60% of the sequences the greater proportion of the sequence was non-coding, the presence of which is likely to have biased the score towards non-coding. This suggests that the actual error rate of non-coding prediction is likely to be lower. It also highlights a limitation in the capacity of the software to handle coding sequences with large untranslated regions.

© University of Pretoria

Chapter 3. Analysis of *Rhipicephalus appendiculatus* Salivary Gland Expressed Sequence Tag Databases (RaGI): Additional Data and Novel Insights

3.1 Overview

The *Rhipicephalus appendiculatus* Gene Index (RaGI), housed at the Dana-Farber Cancer Institute (DFCI) and Harvard School of Public Health, was compiled from mRNA sequences derived from the salivary glands of four-day fed adult female *R. appendiculatus T. parva*-infected and un-infected ticks from the Muguga tick stock maintained at the International Livestock Research Institute (ILRI) (Nene et al., 2004).

Preparation of tick material, EST library construction and the pipeline for assembly and automated annotation of RaGI is outlined in the Methods section (Chapter 2.3). This pipeline has been used to construct gene indices for 60 plant, 45 animal, 15 protozoan and 10 fungal species, all of which are hosted at DFCI. Amongst these are four tick species – *Rhipicephalus appendiculatus*, *Rhipicephalus microplus*, *Amblyomma variegatum* and *Ixodes scapularis*.

Data incorporated within RaGI has been re-assembled three times, in order to incorporate additional sequence data generated since 2004. The updates are designated Release 1.0, Release 2.0 and Release 2.1, the latter being the most current one updated on July 14 2008. Release 1.0 (Nene et al., 2004) comprising 18,422 ESTs, identified 2,543 TCs, and 4,797 unassembled ESTs (singletons). Therefore it comprises 7,340

© University of Pretoria

unique sequences. The latest assembly, Release 2.1, made available in July 2008, includes additional sequences generated by screening for tick salivary gland polypeptides with signal peptides that were functional in heterologous systems (Lambson et al., 2004). This version comprises 18,509 ESTs that are assembled into 2,642 tentative consensus (TC) sequences and 4,917 singletons.

As only 87 ESTs were added to Release 2.1 and the improvement in the assembly of the ESTs was negligible, the work in this thesis is based on Release 1.0 of 12 December 2003. The term RaGI will be used to refer to Release 1.0 hereafter, unless stated otherwise, and the sequence identifiers also refer to those in Release 1.0.

The automated annotation carried out during the construction of RaGI assigned a putative function to 1,338 of the 2,543 of the TCs. In total 4,470 of the RaGI sequences including singletons (60.7%) had no tentative annotations (TAs), that is, they did not have significant matches (E-value < 1e-25) in the GenBank NR database. I carried out a more recent BLAST search against the GenBank NR database, using the same search parameters described in the original publication of RaGI (Nene et al., 2004), which resulted in TAs for 206 TC sequences previously labelled as unannotated (Appendix Table A1.1). This was a result of identification of homologues for many previously unknown and uncharacterised tick proteins in the large amount of novel data that has been generated through the *I. scapularis* genome project (Hill and Wikel, 2005). Approximately 80% of the newly annotated sequences matched predicted proteins encoded within the *I. scapularis* genome. The remaining 957 (38%) RaGI unannotated TCs had no matches in the GenBank NR database.

A second EST library was constructed by collaborators at the National Institutes of Health (NIH) from the salivary gland mRNA material of uninfected *R. appendiculatus* ticks. While RaGI was constructed from EST libraries that had been selected for size by gel fractionation and isolation of transcripts > 2 Kb in length, the library generated at NIH was not size selected. It contained 1,784 EST sequences with an average length of 887 bp. This was more than 200 bp longer than that obtained for RaGI (668 bp in

© University of Pretoria

RaGI)(Nene et al., 2004). 1,371 of the NIH ESTs contained a sequenced polyA/T tail (77%). ESTs were assembled into 121 TCs from 557 ESTs, and 1,081 singletons remained. The dataset contained 1,202 new *R. appendiculatus* transcripts that had average and median lengths of 693 bp and 704 bp, respectively. There was a significant difference (T-test P-value < 0.05) in the mean lengths of the NIH and RaGI assemblies, with the mean length of contigs within RaGI being 858 bp. The longer contigs observed in RaGI is likely due to the less stringent criteria used for the assembly. A BlastX search of the 1,202 sequences resulted in 514 sequences with significant matches (E-value < 1e-25) to proteins in GenBank NR. Of those, 83 *Theileria parva* and mammalian host contaminant sequences were filtered out, leaving 431 (39%) as annotated tick sequences. The proportion of annotated tick sequences is comparable to that observed in RaGI (~40%). About 50% (542 sequences) of the 1,202 NIH sequences are novel as defined by lack of sequence identity with sequences in RaGI. The NIH *R. appendiculatus* sequences are significantly shorter (mean length = 620 bp) than their matches in RaGI (mean length = 760 bp) indicating that small transcripts may be lost if size fractionated libraries are used for EST construction. Of the newly identified RA sequences 124 have sequence matches with an E-value of < 1e-25 to other tick species and insects in the GenBank NR database. A majority of these newly identified annotated sequences putatively encode housekeeping proteins including ribosomal proteins, mitochondrial proteins, metabolic enzymes such as NADH dehydrogenase, ATPases, esterases, and structural proteins such as fibrillin and myosin. A partial homologue (52% of the full length protein; 92% protein identity) of Boophilin, a thrombin-inhibiting protein identified in *R. microplus* (GI:17529564; PDB: 2ODY), was also found amongst the new sequences. The remaining ~400 sequences which have no significant matches (E-value < 1e-25) in the GenBank NR database are discussed in Chapter 5.

The high percentage of novel sequences within this relatively small additional RA dataset generated from the same starting RNA sample indicates that the specific technical approach used is important in determining the composition of EST databases. The proportion of new sequences identified in a single tissue, salivary gland, also raises the issue of what level of coverage is required to define relatively complete

© University of Pretoria

transcriptomes from arthropods with large complex genomes, such as those of ixodid ticks (Guerrero et al., 2006).

3.2 Summary of Gene Families within RaGI

Over 3,400 transcribed sequences from the salivary glands of various tick species were categorized into gene families in a review by Francischetti et al., (2009). Amongst them were 1,985 *R. appendiculatus* sequences. Some of them are presented in Appendix Table A1.2. The RaGI accessions were identified and the functional annotations were examined at individual sequence level here. Francischetti et al., (2009) describe 22 major gene families of which sequences from *R. appendiculatus* fall in 13. They are described below.

3.2.1 Glycine-rich superfamily

Vertebrate and invertebrate extracellular matrix proteins such as collagen, keratin and elastin, as well as silk fibroins have a high glycine content. The absence of a side chain in glycine makes it an important component of fibrous structural proteins, which frequently assemble into triple helical secondary structures.

In ticks, immunogenic glycine-rich proteins have been isolated from the cement cone (Bishop et al., 2002; Trimnell et al., 2002). Within RaGI, 70 peptides were characterized in this category. Nine of these peptides were similar to cuticle proteins of other insects, including mosquitoes and ticks, five were similar to collagen and 11 were similar to spider silk-like peptides. Three peptides were specific to the *Rhipicephalus* genus. However it should be noted the amino acid composition bias means that such similarities should be interpreted with caution.

Glycine-rich proteins in tick EST datasets are discussed in more detail in section 3.5.

© University of Pretoria

3.2.2 Mucins

Mucins are heavily glycosylated proteins. They bind to the chitinous lining of epithelial cells including those present in salivary glands and insect mouthparts. One of their functions in insects is probably to lubricate the mouthparts (Arcà et al., 2005). BMA7, a mucin-like protein from *R. microplus* was partially protective against tick infestation when it was used to immunize cattle (McKenna et al., 1998; de la Fuente et al., 2006c).

Mucins have been identified in the salivary gland transcripts of mosquitoes and ticks (Arcà et al., 2005; Ribeiro et al., 2006; Ribeiro et al., 2007). RaGI contains 18 putative mucin proteins, two of which contain the chitin binding peritrophic domain-A (CDD: cl02629; pfam: pfam01607), ChtBD2. Removal of ChtBD2 from proteins resulted in lower affinity for chitin in insects (Jasrapuria et al., 2010).

3.2.3 Antigen 5 family

Four peptides within RaGI had similarity with a family of proteins designated the antigen-5 family. Proteins in this family are commonly found in the venom of snakes, lizards and wasps (Fang et al., 1988; Hoffman, 2006), in blood feeding insects such as mosquitoes, sand flies and tsetse flies (Francischetti et al., 2002; Kato et al., 2006; Li et al., 2001a; Valenzuela et al., 2002b) as well as in plant defence proteins. A protein belonging to this family was also identified in *I. scapularis* (Ribeiro et al., 2006).

The four peptides exhibited database hits to conserved domain sperm-coating protein (SCP) superfamily (CDD: cl00133). The biological roles of animal proteins in this family are not well known however the helothermine protein produced by the beaded lizard is believed to be a neurotoxin (Nobile et al., 1996).

This constitutes an interesting example of wide phylogenetic conservation of a domain in proteins that probably have distinct functions in different taxa.

© University of Pretoria

3.2.4 PK domain families

Prokineticin (PK)-like cytokine, which contains a PK domain, is involved in haematopoesis and blood cell differentiation in invertebrates (Söderhäll et al., 2005). The better-studied vertebrate PK proteins are thought to be involved in physiological processes such as angiogenesis and neurogenesis.

Amongst vertebrates, the PK domain is also present in Mamba intestinal toxin – 1 (MIT1) protein, which has a protein fold similar to Colipase, an enzyme involved in lipid hydrolysis. A PK-like domain is also found in Bv8, a prokineticin from the toad *Bombina variegata*, that was shown to stimulate contraction of gastrointestinal smooth muscles in guinea pigs (Li et al., 2001b) and rabbits (Lai et al., 2003). Structural similarities to PK domains were also found in Dickkopf (DKK)–related proteins. In vertebrates DKK proteins antagonise Wnt signalling pathway that is involved in embryonic development and normal physiological processes in adult animals. Homologues of vertebrate DKK proteins have not been reported in insects or nematodes however DKK-like proteins have been identified in the venom of spiders (Szeto et al., 2000) .

Three peptides containing PK domains were identified in RaGI by searching against PFAM. All three had amino acid similarity with proteins containing PK domains. Despite having weak sequence identity with each other the cysteine residues that form disulphide bridges were conserved between all three peptides.

3.2.5 Proteins containing protease inhibitor domains

This family contains proteins with Kunitz domains, many of which are serine protease inhibitors (serpins), cysteine protease inhibitors, trypsin inhibitor-like (TIL) and hirudin/madanin-like proteins. An example of a member of this family is the Tick Anticoagulant Peptide (TAP) that inhibits factor Xa in the blood coagulation pathway in the soft tick *Ornithodoros moubata* (Waxman et al., 1990).

© University of Pretoria

Thirty-five RA peptides contain protease inhibitor domains. Amongst these, 12 peptides have similarities with serpins (discussed in detail in section 3.3.2), one is a cysteine protease inhibitor, five contain the TIL domain, three are hirudin-like. It seems probable that many of these are involved in modulation/suppression of the host haemostatic response.

3.2.6 Basic tail and related secreted proteins

This group currently contains eight proteins. Salp14, an anticoagulant that inhibits factor Xa, isolated from *I. scapularis,* is a well characterized member of this family (Narasimhan et al., 2002).

3.2.7 Lipocalin fold superfamily

Proteins in this family vary greatly in sequence but have a conserved structure comprising an eight-stranded, antiparallel, β-barrel enclosing an internal ligand-binding site. This family includes a group of unusual tick-encoded histamine-binding proteins (described in detail in section 3.3.3). Eighteen RaGI sequences have similarity with the lipocalin domain. Five of these have sequence similarity with the histamine-binding superfamily (pfam02098). Structure prediction using GenThreader of the 18 sequences produced structural matches to lipocalin structures with P-values less than 0.074 for 14 peptides previously identified in soft ticks (Table 3.1). This confirms that structural conservation occurs between proteins in phylogenetically very distinct families of ticks.

© University of Pretoria

Table 3.1 GenTHREADER protein structure prediction of R. appendiculatus sequences reveals matches to lipocalin folds of soft ticks.

Sequence Id	Confidence Level[1]	Score[2]	p-value[3]	Epair[4]	Esolv[5]	AlnSc[6]	Alen[7]	Dlen[8]	Tlen[9]	PDB_ID Top Hit
TC1624	CERT	76.617	9.00E-07	-279.6	-12.1	352	126	126	131	3d9yA0
TC994	CERT	71.479	3.00E-06	-210.9	-16.8	311.5	171	175	177	1qftA0
CD791328	MEDIUM	44.317	0.002	-166.8	-10.2	151	153	153	196	3brnA
TC410	MEDIUM	41.736	0.003	-224.9	-5	126	149	153	217	3brnA0
CD782167	MEDIUM	41.648	0.003	-229	-7	118	149	153	214	3brnA0
CD788077	MEDIUM	40.831	0.004	-195.3	-12.7	113	147	153	196	3brnA
TC1235	MEDIUM	39.328	0.005	-96.8	-9.5	137	135	153	199	3brnA
TC2294	MEDIUM	39.011	0.006	-153	-10.9	125	127	144	248	2x46A0
TC1593	MEDIUM	38	0.007	-186.3	-8.4	105	148	153	217	3brnA0
TC2494	LOW	36.417	0.011	-209.9	-8.9	88	138	146	206	3bs2A
TC1138	LOW	35.739	0.012	-105.8	-7.2	113	141	153	192	3brnA
TC1154	LOW	34.686	0.016	-49.3	-8.2	113	143	145	192	2cm4A
TC780	LOW	31.909	0.03	-134.6	-8.6	83	127	144	271	2x46A0
CD795421	LOW	28.088	0.074	-73.1	-6.2	72	150	153	204	3brnA
CD784141	GUESS	25.841	0.125	-142.5	-4.2	53	97	222	144	1n0xH0
TC1143	GUESS	24.768	0.16	-21.1	1.3	85.4	20	20	272	2hwnF0

[1] CERT (p-value < 0.0001); MEDIUM (p-value < 0.01); LOW (p-value < 0.1); GUESS (p-value ≥ 0.1); [2] Score: raw score from support vector machine; [3] p-value: Probability of false positive; [4] Epair: Pairwise energy for model; [5] Esolv: Solvation energy for model; [6] AlnSc: Sequence alignment score; [7] Alen: Alignment length; [8] Dlen: Length of PDB entry; [9] Tlen: Length of target sequence

© University of Pretoria

3.2.8 Ixodid 8.9 kDa peptide family

Seven short predicted peptides, 8.9 kDa in size, have similarities to secreted proteins identified in other Ixodid ticks - *R. sanguineus* (gi: 260908312), *Hyalomma marginatum rufipes* (gi: 307006471), *Amblyomma maculatum* (gi: 346471731), *I. scapularis* (gi: 67083146). No domains conserved outside the I*xodidae* were identified within these peptides.

3.2.9 Enzymes: proteases, nucleases, esterases, lipases, chitinases

Enzymes including metalloproteases, nucleotidase/apyrase, carboxypeptidase, chitinase, serine proteases, carboxyl esterase, endonucleases and phospholipase are represented by 80 proteins, 37 of which are metalloproteases, as identified by conserved domain matches to the Zn-dependent metalloproteases secreted by arthropod salivary glands (CDD: cd04272, Superfamily: cl00064) (Francischetti et al., 2003). Most of these contain the Pfam reprolysin motif (PF01421). Metalloproteases have been found to be expressed abundantly in other hard ticks (Chmelař et al., 2008; Nakajima et al., 2005; Ribeiro et al., 2006; Valenzuela et al., 2002a) as well as soft ticks (Mans et al., 2008a) and are thought to be involved in anti-blood clotting activity (Valenzuela et al., 2002a).

3.2.10 Host immune and inflammatory response antagonists

Tick salivary glands are known to secrete proteins that modulate the host response to tick feeding. In RaGI 13 sequences were identified in this category. Amongst them are antimicrobial peptides, such as defensins and histidine-rich proteins that could be preventing microbial growth and disruption of the tick feeding (Ribeiro et al. 2006). The protease inhibitor alpha-2-macroglobulin (two singletons and one TC in RaGI) is also likely to be involved in host interaction. Ten of these sequences contain a predicted signal peptide at the N-terminal end of the ORF. These include a galactoside-binding lectin, ML (MD-2-related lipid-recognition) domain containing protein, alpha-2-

© University of Pretoria

macroglobulin, defensin, microplusin peptide and several histidine-rich proteins. Twenty-one ESTs were assembled into seven TCs, all containing low numbers of transcripts (< 5 ESTs), while six were singletons. All but two of the TCs (TC1081, TC1145) were assembled from ESTs derived from both infected and uninfected libraries. TC1081, comprising two ESTs from the infected library, contains an ORF 134 aa long that has structural similarity (HHPred probability score = 97.95; SMART e-value = 0.85) with the human Ly6 neurotoxin-like protein 1 (LYNX1), a nicotinic acetylcholine receptor binding protein (PDB:2L03; PFAM: PF00021; smart00134) that is also present in other mammals. The Ly-6 domain is located between residues 23 and 73 of the ORF. The Ly-6 domain has not previously been reported in ticks however literature search on domains revealed that several mosquito sequences have been found to contain it.

The 178 aa ORF contained within TC1145, which was assembled from two ESTs from the uninfected library, had sequence similarity with the ML (MD-2-related lipid-recognition) domain, which is involved in interaction with lipids that are important in innate immunity. The ML domain has been reported in the hard ticks *I. scapularis*, *I. ricinus* and *D. variabilis* as well as in soft ticks *Ornithodoros parkeri* and *O. coriaceus*. The ML domain was also detected in eight *R. microplus* sequences by sequence search against the Interpro database.

3.2.11 *R. appendiculatus* Orphan Proteins

491 RA EST sequences had no protein sequence-based matches with known families. RaGI, derived from *R. appendiculatus* salivary glands has by far the highest proportion of orphan proteins (25%) as compared to EST databases derived from other tick species (Table 3.2). This could be a consequence of a non-normalized library, which RaGI as well as AvGI (22% orphan proteins) are. Orphan proteins are investigated further in Chapter 4.

© University of Pretoria

Table 3.2 Number of orphan proteins identified by sequence similarity methods in hard and soft ticks in Francischetti et al., 2009.

Tick Species	# Total Sequences in Table	# Predicted Orphan Proteins	% Orphan sequences
Amblyomma americanum	397	33	8%
Amblyomma cajennense	230	44	19%
Amblyomma variegatum	396	86	22%
Dermacentor andersoni	363	52	14%
Haemaphysalis	130	3	2%
Rhipicephalus appendiculatus	1985	491	25%
Rhipicephalus microplus	2336	208	9%
Ixodes	1087	34	3%
Argas monolakensis	196	16	8%
Ornithodoros parkeri	168	15	9%

3.2.12 Conserved secretory pathway proteins, with putative housekeeping functions by analogy with homologues in other taxa.

The 18 peptides classified in this group all had a predicted signal peptide. These proteins include salivary gland-encoded selenoproteins that are thought to be involved in the metabolism of oxidising compounds (Beckett and Arthur, 2005), calreticulins that are responsible for quality control of folded proteins in the endoplasmic reticulum, together with other conserved secretory pathway proteins. The majority of the proteins in this group could be involved in the mechanics of the secretory pathway.

3.2.13 Non-secreted conserved proteins.

Sequences in this category are primarily housekeeping genes including transcriptional and translational regulatory proteins, ribosomal proteins, cytochromes, heat shock proteins, histones and mitochondrial proteins. This category contains 961 sequences, and constitutes a majority of the annotated sequences in RaGI.

© University of Pretoria

3.3 Homologues of tick genes encoding previously identified vaccine candidates

RaGI protein homologues of several tick proteins that were characterized in earlier studies, were identified. These include RaGI homologues of the 20/24 kDa immuno-dominant protein from *R. appendiculatus* salivary glands (AY208825), a p36 T cell inhibitor from *Dermacentor andersoni* (Bergman et al., 2000) and the *R. microplus* heme-lipoproteins A and B (Maya-Monteiro et al., 2000). A preliminary description of these is provided in the initial publication of RaGI (Nene et al., 2004).

RaGI homologues of five additional previously characterized tick vaccine candidates, including proteins from *R. microplus, H. longicornis* and *I. ricinus*, were identified in this work, namely RIM36, Serpins, Histamine Binding Proteins, Subolesin and Bm91.

3.3.1 RIM36

RIM36 is a 36 kDa immuno-dominant molecule isolated from the salivary glands of *R. appendiculatus* (Bishop et al., 2002). It is a component of the tick cement and is expressed in the type III acini cells of the salivary glands. The protein is 334 residues long, contains a predicted signal peptide and is glycine-rich.

RIM36 has three variants in the RaGI database. TC1398 is 100% identical to AAK98794, the protein described in Bishop et al., 2002. TC1399 and TC1400 are 97% and 96% identical at the protein level with AAK98794, respectively. Figure 3.1 shows a multiple alignment, generated using Muscle (Edgar, 2004) alignment algorithm, of the variants of the RIM36 sequences. TC1400 has one less copy of the tri-peptide GL[G/Y/S/F/L] and one more copy of the septa-peptide GSPLSGF while TC1399 has two fewer copies of the latter motif compared to RIM36 and TC1398. Additionally, the TC1399 variant is lacking 31 residues located at the C-terminus. These multiple variants emphasise the complexity and redundancy of the *R. appendiculatus* sialome. This variation may have

© University of Pretoria

implications for the use of RIM36 for vaccination and could be a factor contributing to the lack of protection induced by recombinant RIM36 in immunised rabbits and cattle (R. Bishop and A. Musoke, unpublished data).

```
RIM36           MKVFVAVALISLGIHAEAFYGGSFSSGFGSGLYGGGLFGSGLSGFYGGGLGSGLYGGGLL
TC1398          MKVFVAVALISLGIHAEAFYGGSFSSGFGSGLYGGGLFGSGLSGFYGGGLGSGLYGGGLL
TC1400          MKVFVAVALISLGIHAEAFYGGSFSSGFGSGLYGGGLFGSGLSGFYGGGLGSGLYGGGLL
TC1399          MKVFVAVALISLGIHAEAFYGGSFSSGFGSGLYGGGLFGSGLSGFYGGGLGSGLYGGGLL
                ************************************************************

RIM36           GGGLSGLSGLYGGGLYGLGSTRVLGLGGLGGLGGLGGLGGLSGFGYPGLFGSRYGGLLGG
TC1398          GGGLSGLSGLYGGGLYGLGSTRVLGLGGLGGLGGLGGLGGLSGFGYPGLFGSRYGGLLGG
TC1400          GGGLSGLSGLYGGGLYGLGSTRVLGLGGLGGLGGLGGL---SGFGYPGLFGSRYGGLLGG
TC1399          GGGLSGLSGLYGGGLYGLGSTRVLGLGGLGGLGGLGGLGGLSGFGYPGLFGSRYGGLLGG
                ************************************   ******************

RIM36           YGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSP
TC1398          YGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSP
TC1400          YGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSP
TC1399          YGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSP
                ************************************************************

RIM36           LSGFGSPLSGFGSPLSGFGSPLSGFGSP-------FGSYGPLSMGLGAPRRFPGDLRLIS
TC1398          LSGFGSPLSGFGSPLSGFGSPLSGFGSP-------FGSYGPLSMGLGAPRRFPGDLRLIS
TC1400          LSGFGSPLSGFGSPLSGFGSPLSGFGSPLSGFGSPFGSYGPLSMGLGAPRRFPGDLRLIS
TC1399          LSGFGSPLSGFGSPLSGFGSP--------------FGSYGPLSMGLGAPRRFPGDLRLIS
                *********************              ************************

RIM36           EPTSRLPVSDAVYIGVVRQRLVPAPYRVPHPVPVPQPFPVPQPYQVDVPVPKPVEVPVPR
TC1398          EPTSRLPVSDAVYIGVVRQRLVPAPYRVPHPVPVPQPFPVPQPYQVDVPVPKPVEVPVPR
TC1400          EPTSRLPVSDAVYIGVVRQRLVPAPYRVPHPVPVPQPFPVPQPYQVDVPVPKPVEVPVPK
TC1399          EPTSRLPVSDAVYIGVVRQRLVPAPYRVPHPVPVPQPFPVPQPYQVDVPVPKAVEVPVPK
                ************************************************.******.

RIM36           PEPIHTVTEVTNTAVVQPVTSAVVHTEGPHVTGQVQEHVAI
TC1398          PEPIHTVTEVTNTAVVQPVTSAVVHTEGPHVTGQVQEHVAI
TC1400          PEPIHTVTEVTNTAVVQPVTSTVVHTEGPHVTGQVQEHVAI
TC1399          PEPIHTVTEV------------------------------
                **********
```

Figure 3.1 Multiple protein alignment of the R. appendiculatus glycine-rich protein, RIM36 and its variants identified in R. appendiculatus Gene Index– TC1398, TC1399 and TC1400. The signal peptide is underlined. The stars indicate conservation of the residues at that position.

3.3.2 Serine Protease Inhibitors (Serpins)

Serpins are the largest and most diverse family of protease inhibitors currently identified in biological systems. Most serpins inhibit proteases and regulate proteolytic pathways but some have other functions and are involved in processes such as storage, hormone transport, and tumour suppression. The role of serpins in modulating coagulation and inflammatory pathways has led some researches to believe that they regulate haemolymph coagulation in ticks (Maritz-Olivier et al., 2007).

© University of Pretoria

Recombinant serpins from several tick species have been shown to induce host immune responses following experimental immunisation and therefore represent vaccine candidates (Andreotti et al., 2002; Imamura et al., 2005; Leboulle et al., 2002; Prevot et al., 2007; Sugino et al., 2003). This includes those from *R. appendiculatus* (Imamura et al., 2006).

Messenger RNAs (mRNAs) encoding four serpins have been isolated from the midgut of *R. appendiculatus* – Serpin-1 to 4 - GB accessions AAK61375–8, respectively (Mulenga et al., 2003). In the RaGI database TC1759 and TC1592 are full-length homologues of Serpin-2 and 3, respectively. However, both these homologues contain differences comprising insertion/deletions (indels) and single nucleotide polymorphisms (SNPs) relative to the gut *R. appendiculatus* transcripts. In the case of TC1759 these polymorphisms result in a putative frameshift in the coding sequence. TC1759 is composed of two ESTs, which are mate-pairs from a single clone. The sequence polymorphisms observed between mRNA AAK61376.1 and TC1759 may be due to sequencing errors. This can be verified by deeper sequence coverage. This comparative analysis and two possible interpretations again highlight the issue of what represents adequate coverage in EST dataset, this time from the perspective of accuracy in relatively rare transcripts.

TC1592, a full-length homologue, has 96% amino acid identity with Serpin-3. It is possible that this reflects genetic drift in the sequences between the *R. appendiculatus* stocks used in the two different studies. The two sites that are thought to be conserved from mammalian to arthropod serpins (Mulenga et al., 2001) are present in both TC1759 and TC1592, consistent with their functionality. There are no homologues of Serpin-1 in RaGI; Serpin-4 has partial sequence similarity in TC1528. The amino acid sequence identity is 72% over the conserved sections of the protein but the predicted salivary gland protein is truncated 150 residues from the carboxyl terminus, as a consequence of the occurrence of indels between the two sequences. The observed difference may be a result either of genetic polymorphism between the different tick

© University of Pretoria

stocks used as source material, or of expression of distinct variants in the two different tissues - gut and salivary glands, respectively, perhaps related to distinct functions.

Keyword search for “serpin” and “lospin” in the gene indices resulted in eight matches in RaGI (4 TCs and 1 EST), 23 matches in BmiGI (17 TCs and 6 ESTs), seven matches in AvGI (1 TC and 6 ESTs) and 89 matches in IsGI (56 TCs and 33 ESTs). Forty-five genes encoding serpins were identified through sequence homology in the *I. scapularis* genome (Mulenga et al., 2009). Differential expression of these genes was observed between midgut and salivary glands as well as between unfed and partially fed ticks (Mulenga et al., 2003).

It can be concluded that serpins are highly expressed components of several ixodid tick tissues and that, although best characterised in the gut of *R. appendiculatus* (Mulenga et al., 2003) they are also expressed at the RNA level in salivary glands. Proteomics analyses will be required to ascertain whether there is also expression of serpin proteins in salivary glands. If this proves to be the case, it seems likely that the functions will be different and may involve interaction with host pathways.

3.3.3 Histamine Binding Proteins (HBP)

Tissue damage caused by the tick bite initiates an inflammatory response from the mammalian host. One component in this is the release of the molecule histamine that binds to receptors H1 and H2 on the surface of target cells. Anti-inflammatory drugs targeting the histamine pathway work by competing with histamine for binding to the H1 or H2 receptors. The mechanism whereby ticks antagonise the histamine release pathway, however, involves direct binding of the tick protein to the histamine molecule (Paesen et al., 2000). Production of histamine, which is released mainly by mast cells that are a component of the innate immune pathway, is likely an ancient mechanism for controlling multicellular parasites. It is primarily localised in the gastrointestinal tract and the skin (Nuttall and Labuda, 2004). Other blood-feeding arthropods also produce molecules that bind directly to histamine, blocking it from reaching the target cells. One

© University of Pretoria

example is the nitrophorin family of *Rodnius prolixus* (a vector of *Trypanosoma cruzi* that causes Chaga's disease).

Three HBPs were previously isolated from the salivary glands of *R. appendiculatus* using affinity chromatography (Paesen et al., 1999); Ra-HBP1 and Ra-HBP2 from female ticks and Ra-HBP3 from male ticks. The protein structure of Ra-HBP2 has been solved by X-ray crystallography and shown to be part of the widespread family of proteins that possess the lipocalin fold. Unusually for arthropod encoded lipocalins it possesses two ligand-binding sites – L and H with distinct histamine binding affinities. Ra-HBP1 and 2 have a net charge of -21, which gives the molecular surface of the proteins a strong electrostatic potential that probably contributes to recruitment of cationic compounds including histamine, prior to binding.

The two female-specific Ra-HBPs have higher overall sequence identity to each other than to the male-specific one. However, the L pockets of Ra-HBP2 and 3 are more similar to each other than to that of Ra-HBP1. There is only one difference in the H sites of Ra-HBP1 and 2. But it appears the differences in the L-site have more influence over the overall affinity of the protein for binding to histamine, therefore HBP2 and HBP3 may play the dominant role in binding histamine, despite being encoded by female and male ticks, respectively.

The female-specific HBPs were not detected in male, and only in adult females, not nymphs or in larvae, while the male-specific HBP was detected in adults, nymphs and larvae but not in any instar of females. This is consistent with a role of males where they promote feeding in females at all developmental stages (Wang et al., 1998).

Females secrete HBPs in the early periods of feeding, peaking at ~48hrs after infestation (Paesen et al., 2000). Males produce HBPs throughout their feeding cycle as they attach to the host several times. Female *I. ricinus* exhibited a different pattern of expression where HBPs were absent from a library derived from unfed ticks, however, they were observed in 24 hrs to 7-day fed ticks, increasing in quantity with time

© University of Pretoria

(Chmelař et al., 2008). HBP1 is not very efficient in binding histamine (Paesen et al., 2000) but it is secreted at the same time as HBP2. Therefore, it might have an alternative anti-inflammatory function.

Monotonin and Monomine are two abundant proteins containing the widespread lipocalin fold that binds to histamine in *Argas monolakensis*, a soft tick (Mans et al., 2008c). Both lipocalins have a single binding site (instead of the two observed in Ra-HBPs). The binding site is similar to the L site of Ra-HBPs. The differences in histamine binding protein structure between the soft ticks (Argasidae) and hard ticks (Ixodidae) suggests that the histamine binding function may have evolved after the split of the two lineages. However, it is also possible that one site may have been lost, post divergence.

One homologue corresponding to Ra-HBP2 could be identified in RaGI on the basis of sequence identity. TC994 consisting of two ESTs has 97% protein sequence identity with the full length Ra-HBP2 (Figure 3.2). The four Cysteine and four Tryptophan residues, along with all the other key functional residues in the L and H binding sites are conserved (illustrated in bold in Figure 3.2).

```
TC994             1 MKLLILSLALVLAFSQVRGNQPDWADEAANGAHQDAWKSLKADVENVYYM  50
                    |||||||||||||.|||:|||||||||||||||||||||||||||||||||
AAC63107.1        1 MKLLILSLALVLALSQVKGNQPDWADEAANGAHQDAWKSLKADVENVYYM  50

TC994            51 VKATYKNDPVWGNDFTCVGVMANDVNEDEKSIQAGFLFMNNADTNMQFAT 100
                    ||||||||||||||||||||||||||||||||.|||||||||||||||||
AAC63107.1       51 VKATYKNDPVWGNDFTCVGVMANDVNEDEKSIQAEFLFMNNADTNMQFAT 100

TC994           101 EKVTAVKMYGYNRENAFRYETEDGQVFTDVIAYSDDNCDVIYVPGTDGNE 150
                    ||||||||||||||||||||||||||||||||||||||||||||||||||
AAC63107.1      101 EKVTAVKMYGYNRENAFRYETEDGQVFTDVIAYSDDNCDVIYVPGTDGNE 150

TC994           151 EGYELWTTDYENIPASCLNKFNEYAVGRETRDVFTSACLE 190
                    ||||||||||:||||:|||||||||||||||||||||||
AAC63107.1      151 EGYELWTTDYDNIPANCLNKFNEYAVGRETRDVFTSACLE 190
```

Figure 3.2 A global pairwise alignment of R. appendiculatus female-specific histamine-binding protein 2 (AAC63107.1) with TC994 from RaGI. The conserved Cysteine (C) and Tryptophan (W) residues are in bold. Vertical lines indicate identical residues and dots indicate chemically similar residues.

© University of Pretoria

The scarcity of ESTs coding for HBPs in RaGI is not entirely surprising as the RaGI library is constructed from predominantly female ticks that were fed for four days. Expression of HBPs, which peaks at about 48 hrs (two days) and should therefore have diminished by the fourth day. This highlights a limitation of single time-point EST datasets, which could be addressed by obtaining data from multiple time points and reducing the complexity of the input material by using subtractive methods as was done for BmiGI (Guerrero et al., 2005). However, using the latter approach quantitative information will be lost. A third approach to EST library construction is to pool data from all the tissues. However, both quantitative as well as tissue-specific information is missed using this method.

3.3.4 Subolesin

Subolesin, a protein that was identified using RNAi is thought to be involved in modulation of tick feeding and reproduction, and is a protective antigen that reduces *I. scapularis* infestation in a murine model (Almazán et al., 2003; de la Fuente et al., 2006a). It was shown to reduce infestation against all tick developmental stages. Heterologous protection induced by *I. scapularis* recombinant subolesin was also observed against challenge by *D. variabilis* and *A. americanum* species. Subolesin is conserved not only in several tick species (Nijhof et al., 2007; de la Fuente et al., 2006a), but in nematodes as well as in vertebrates, including humans. This, together with its potential to cross-protect makes it a strong vaccine candidate that could be universally effective for broad-spectrum control of Ixodid ticks.

One homologue of *I. scapularis* subolesin 4D8 mRNA, isolated from *R. appendiculatus,* GI:77166555 (de la Fuente, 2006, unpublished), was also found in RaGI. TC107 encodes the entire 161 amino acids with 100% sequence identity to GI:77166555 at the nucleotide level. TC107 has 86% nucleotide and 98% protein identity with *R. microplus* subolesin, ABI79458 (Nijhof et al., 2007) and 95% nucleotide and 98% protein identity with the *R. sanguineus* homologue (DQ159968) (de la Fuente et al., 2006b). As already mentioned, the high level of conservation between subolesins from various ixodid tick

© University of Pretoria

species makes it a strong vaccine candidate that should be evaluated for control of *R. appendiculatus* infestation.

3.3.5 Bm91

Bm91, a carboxydipeptidase, expressed largely in the salivary glands and the midgut, is a concealed antigen that has been shown to increase the efficacy of the Bm86 vaccine when co-administered (Willadsen et al., 1996).

In RaGI TC400 presents a 3' truncated homologue of Bm91. It is 92% identical over the first 273 residues of Bm91. The absence of the corresponding sequence for the remaining 387 aa is likely to be due to insufficient sequence coverage over that region. Given its ability to enhance Bm86 efficacy it would be worth isolating the full length *R. appendiculatus* protein. This could be achieved by screening salivary gland cDNA using the 3' RACE technique (Lambson et al., 2005 and references therein).

3.4 Unannotated Transcripts in RaGI

A total of 4,221 *R. appendiculatus* TCs and singletons are currently lacking significant matches with sequences in the public databases and can therefore be considered unannotated. The properties of the TCs amongst these, in terms of length, GC content, number of ESTs consisting the TC and longest ORF are presented in Appendix Table A1.3. To derive a clearer understanding of the nature of putative protein coding sequences in *R. appendiculatus* I first predicted the Open Reading Frames (ORFs) present within the unannotated sequences using Artemis annotation tool. 575 (47.7%), 362 (30%) and 2 (0.17%) unannotated TCs contain no ORFs with amino acid length of more than 120 (360 bp), 100 (300 bp) and 50 (150 bp), respectively.

Given the combination of the lack of significant homology to genes in the non-redundant database, presence of polyA tails implied by the method used for library construction and confirmed empirically for selected sequences (Section 6.2.1; Sunter et al., 2008),

© University of Pretoria

and the absence of ORFs longer than 100 amino acids – characteristics typical of non-coding RNA (ncRNA) - it seems very likely that a proportion of the unannotated sequences represent ncRNA. ncRNAs are discussed in detail in Chapter 4.

Studies in the model arthropod *Drosophila* have described several gene models that do not conform to the conventional 'one gene-one transcript' hypothesis. One model suggests that genes could be nested within other genes. For example, a gene could be encoded in the intron of another gene. It has been observed that 7.5% of genes in the *D. melanogaster* euchromatin are nested inside another gene (Misra et al., 2002). It was also noted that most of these genes were transcribed from the antisense strand as defined by the orientation of the longer 'host' gene in which it was located. Misra et al. also describe additional gene models where non-protein coding genomic sequences (untranslated regions – UTRs) are shared between neighbouring genes (overlapping genes), or where two distinct, non-overlapping coding regions are located on a single mRNA (dicistronic genes), or where genes have alternative transcription start points. In addition approximately 20% of genes in *D. melanogaster* are estimated to be alternatively spliced. There is also evidence of occurrence of alternatively spliced transcripts in *Anopheles gambiae* EST datasets (Arcà et al., 2005; Li et al., 2010). The glucose 6-phosphate dehydrogenase gene in the tick *R. microplus* (Genbank Accession: DQ118973) was found to have three variants arising from alternative splicing (unpublished). In *A. gambiae,* 41% of the structural variations observed in transcribed sequences were due to intron retention. Thirty-five percent were a result of alternative donor/acceptor splice sites and 6% were observed to be a result of exon skipping (Li et al., 2010). Similar patterns have been observed in *Drosophila* (Li et al., 2010; Nagasaki et al., 2005), which are very different to the patterns of different types of alternative splicing observed in mammals where the most abundant variation type was exon skipping (Li et al., 2010; Modrek et al., 2001).

Transcribed pseudogenes have also been reported in vertebrate and arthropod genomes. Yao et al. (2006) found that approximately 4% of vertebrate genes were transcribed pseudogenes. Transcribed pseudogenes have also been reported in the EST

© University of Pretoria

dataset of the bumblebee, *Bombus terrestris* (Sadd et al., 2010). Putative pseudogenes identified in RaGI are discussed in Chapter 4.

The subsequent section attempts to elucidate gene models present in *R. appendiculatus* by analysing clusters of sequences that show very high sequence identity to each other yet are assembled as separate contigs in RaGI.

3.4.1 Redundant sequences within RaGI

Numerous RaGI TCs have been identified as being almost identical to other TCs within RaGI over a part of their length. TCs with >90% DNA sequence similarity over >100 bp, and of E-value < 1e-60 were grouped together. In RaGI, these closely related transcript families remain classified as separate clusters, since polymorphisms within the overlapping regions meant that they did not meet the strict clustering criteria defined in the TIGR EST assembly algorithm which clustered sequences that are more than 95% identical over a region of more than 40 nucleotides, in combination with a less than 20 base mismatch at either end into a single contig (Quackenbush et al., 2001). Nonetheless, the similarities suggest that these TCs that are defined as separate within RaGI may be functionally related.

Members of the clusters TC1313-7, TC1345-7 and TC9-11 have sequence identity to known protein-coding genes and some members of the clusters are subsequently tentatively annotated as such. However, members of clusters TC3-5, TC1286-93 and TC1324-6 had no tentatively annotated members. The former three clusters and cluster TC1324-6 are described below while the remaining two clusters are discussed in chapter 4.

© University of Pretoria

Table 3.3 Six groups that comprise highly similar (E-value < 1e-60) unannotated RaGI sequences.

Cluster	TC#	# ESTs	Length (bp)	%GC	TA
TC1313-7	TC1313	14	995	49.0	None
	TC1314	6	975	49.1	None
	TC1315	9	1623	52.4	[1]similar to Pyruvate kinase, partial (30%)
	TC1316	33	1873	53.3	[2]weakly similar to Pyruvate kinase, partial (83%)
	TC1317	7	786	44.8	None
TC1345-7	TC1345	13	1999	51.1	[3]histone H3.3, complete
	TC1346	6	912	59.9	[3]histone H3.3, complete
	TC1347	6	1398	43.1	None
	TC949	2	538	55.4	[4]histone H3, partial (70%)
TC3-5	TC3	93	3236	41.0	None
	TC4	28	2495	37.5	None
	TC5	49	2622	47.6	None
TC1286-93	TC1286	38	1912	40.5	None
	TC1287	3	622	38.7	None
	TC1288	9	1038	39.0	None
	TC1289	13	2323	52.4	None
	TC1291	9	1587	56.6	None
	TC1293	7	2016	40.6	None
	CD783337	1	698	40.1	None
TC1425-6	TC1425	7	1871	42.7	None
	TC1426	3	987	45.2	None
	CD785259	1	744	38.98	None
TC9-11	TC9	26	1626	46.4	None
	TC10	29	4496	52.2	[5]similar to Na+/K+ ATPase alpha subunit, complete
	TC11	70	3985	43.5	None
TC1324-6	TC1324	33	3656	49.2	None
	TC1325	9	982	48.1	None
	TC1326	3	929	50.6	None

[1] UniRef100_Q0KHB6: *Crassostrea gigas* (Pacific oyster)
[2] UniRef100_Q4S1B0: *Tetraodon nigroviridis* (Green puffer)
[3] Human
[4] *Drosophila teissieri*
[5] *Callinectes sapidus*

© University of Pretoria

3.4.1.2 TC1313-7 cluster

This cluster constitutes TCs 1313, 1314, 1315, 1316 and 1317. Details of their TAs and EST numbers are given in Table 3.3. TC1315 and TC1316 are annotated as 'similar to pyruvate kinase', an enzyme that catalyses the final step of glycolysis, while TC1313, TC1314 and TC1317 do not have significant sequence similarity to a known protein, and are thus considered unannotated (refer to Methods section for annotation criteria). Within this cluster, TC1316 contains the highest number of ESTs with 33 and also has the highest G+C content at 53.3% while TC1317 has the least number of ESTs at 7 and the lowest G+C content at 44.8%.

The relationship of each TC in this group to pyruvate kinase is shown as an alignment in Figure 3.3. TC1316 encodes almost the full-length pyruvate kinase (98% coverage) as compared with *Xenopus laevis* (sp|Q92122.1) while TC1315 aligns with 100% identity with the last 206 residues of TC1316 (as well as the pyruvate kinase ORF) at the 3' end in reverse orientation. The 5' ~1000 bp of TC1315 align with TCs 1313, 1314 and 1317, which have no significant matches in the GenBank NR database. The 5' end of TC1313 and TC1314 have sequence identity of 100% with 69 bp (13 residues from 3' end) of the pyruvate kinase ORF. The short length of alignment of TC1313 and TC1314 to pyruvate kinase protein results in high BlastX E-values, which don't meet the annotation criteria.

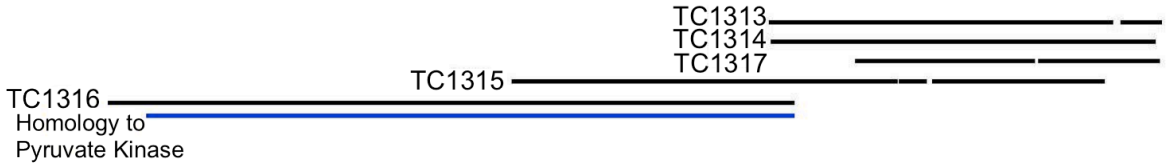


Figure 3.3 Diagrammatic representation of how sequences in cluster TC1313-1317 align relative to each other. Black lines represent aligned nucleotides with breaks indicating alignment gaps. Blue line represents region homologous to pyruvate kinase gene of Xenopus laevis. Figure is to scale.

The longest ORF downstream of the section of the sequence that is homologous to pyruvate kinase is 115 and 134 aa in TC1313 and TC1314, respectively. These two ORFs, although homologous to each other (98% nucleotide identity), do not have significant matches in the GenBank NR database. Sequence downstream of these ORFs

© University of Pretoria

(575 bp in TC1313, 557 bp in TC1314) contains no ORFs longer than 100 aa. The average G+C content decreases considerably from 53% in the pyruvate kinase homologous region to 49% in the region downstream of the latter domain. The ORFs in TC1313 and TC1314 mentioned above have a G+C content of 51% and 54%, respectively.

BmiGI lacks a full-length copy of pyruvate kinase, however partial homologues (BmiGI TC18635, which is composed of four ESTs and CK184874) are present. They have sequence identity above 90% with all the members of the TC1313-7 group (except for TC1317, which aligns with the cluster further downstream) over the pyruvate kinase-coding region only. This suggests substantial divergence from *R. appendiculatus* in the non-coding section of pyruvate kinase transcripts.

A similar cluster of pyruvate kinase-like sequences is also present in IsGI. Eleven exons are annotated for the pyruvate kinase gene in *I. scapularis* (ISCW020197-RA). Exon counts for this gene vary between species. In *X. laevis* the pyruvate kinase gene is encoded by 10 exons. Alignment of transcripts to *I. scapularis* genome assembly shows that the *R. appendiculatus* pyruvate kinase gene is also encoded by 10 exons (Figure 3.4). Orthologues of ISCW020197 gene identified in Vectorbase (release VB-2011-12 - December 2012) show that *A. aegypti*, *A. gambiae* and the human hair louse, *Pediculus humanus* orthologues of pyruvate kinase have four, three and nine exons, respectively. The *I. scapularis* exonic regions are covered by two IsGI TCs - TC45726 that comprises three ESTs and covers exons 1-8; and TC46783 that also contains three ESTs and covers exons 8-11. These two TCs overlap over ~400 bp of exon 8 with 93% nucleotide identity and the predicted translated protein is 98% identical. A third IsGI TC, TC41396, which comprises 8 ESTS, has 98% sequence identity with ~1200 bp of *I. scapularis* sequence that has been annotated as intron 10 of the pyruvate kinase gene in Vectorbase (indicated by green lines in Figure 3.4). This sequence matches 135 bp from the 3' end of exon 10 in *I. scapularis* and extends into the intron. The 135 bp that match exon 10 encode 45 residues of the protein located 58 residues from the 3' terminus of exon 10. By analogy with *R. appendiculatus*, IsGI TC41396 exhibits premature

© University of Pretoria

truncation of the ORF 12 codons from the intron-exon boundary. Sequence searching *I. scapularis* ESTs twelve additional ESTs were found to span the putative intronic region of the *I. scapularis* pyruvate kinase gene. The high level of transcription across the intron strongly suggests that this is not an artefact.

Two isoforms of pyruvate kinase resulting from alternative splicing have been identified in humans (Uniprot: P14618 and P14618-2). Transcript variation in human and mouse genes has also been shown to result from alternative polyadenylation (polyA) sites (Lee et al., 2008) which could be located in the 3'-most exons and introns of the gene (Wang et al., 2009). Alternative splicing to create structural and functional diversity has been demonstrated in *Drosophila* (Tan et al., 2002). Alternative polyadenylation can alter the protein product or the UTR region. There is evidence to suggest that excised introns are processed into regulatory RNA rather than being degraded (Mattick and Makunin, 2006). Assuming that the *I. scapularis* gene is accurately annotated in Vectorbase, it seems quite likely that TC41396 may encode an alternative 3' UTR of the pyruvate kinase gene and also a regulatory RNA. Functional experiments using RNAi, that are now routine in ticks (de la Fuente and Kocan, 2006) would need to be performed to confirm this. As is the case in *I. scapularis*, it is likely that the 28 ESTs contained within RaGI TC1313, TC1314 and TC1317 represent transcription variants.

© University of Pretoria

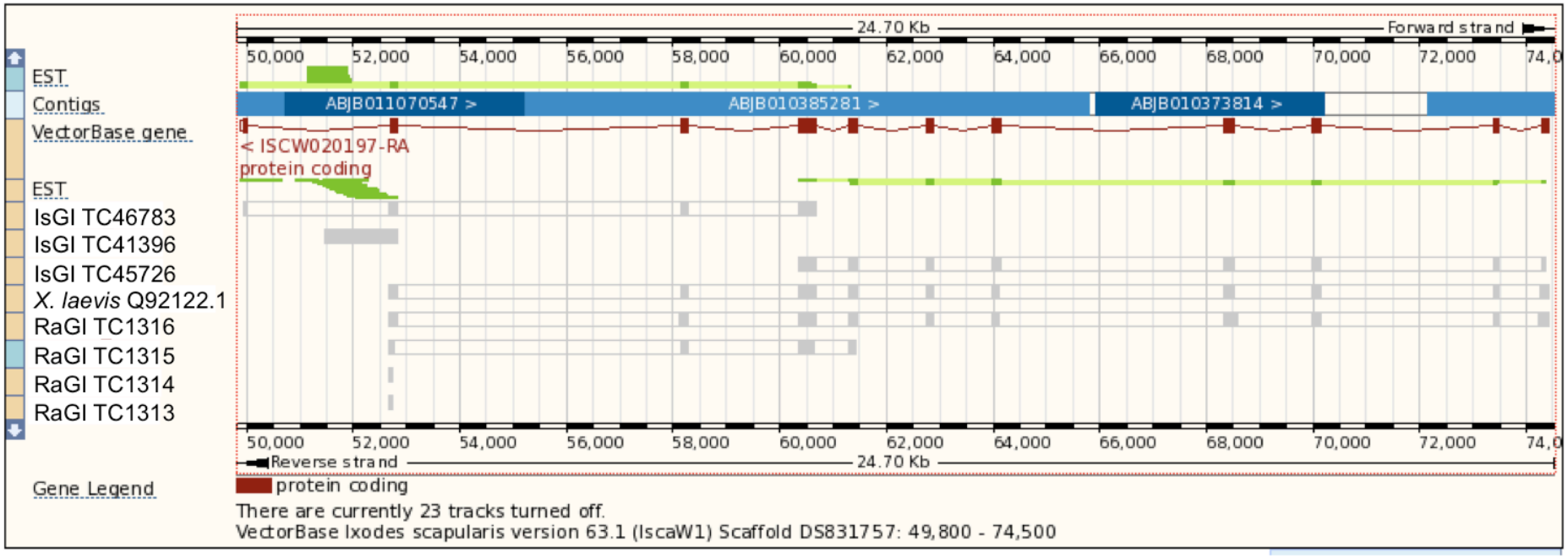


Figure 3.4 ***Presentation of TCs in RaGI cluster TC1313-1317, its IsGI homologues TC46783, TC41396 and TC45726, and X. laevis homologue Q92122.1 (grey bars) onto I. scapularis genome assembly (version 63.1) scaffold DS831757 (location 49,800-74,500 bp). The I. scapularis pyruvate kinase gene, ISCW020197-RA (red line), is encoded on the reverse strand. Red boxes on the gene represent exons while red lines between them represent introns. Shaded grey boxes indicate homology between ISCW020197-RA and the TCs of RaGI and IsGI, and X. laevis Q92122.1. I. scapularis ESTs included in the genome assembly are shown with green lines. ESTs at the 3' end of the gene (dark green lines) show transcription of intron 10. Light and dark blue bars represent the contigs that are assembled into scaffold DS831757 at ISCW020197-RA gene. Figure generated in Vectorbase genome browser.***

© University of Pretoria

3.4.1.3 TC1345-7 cluster

Three TCs are present within this cluster. TC1345 and TC1346 are annotated as Histone H3 and encode the full-length mRNA, ~400 bp, of the variant Histone H3.3, having 100% amino acid identity with the corresponding protein in humans. The coding ORF and the remaining ~300 bp downstream region in TC1346 are 100% and 99% conserved, respectively, in TC1345 and TC1346, however the ~150 bp sequence upstream of the ORF is divergent between the two. TC1347 lacks a full-length ORF encoding Histone H3, however, the partial ORF of 26 aa from the 3' terminus has 100% identity with ORFs in TC1345 and TC1346.

Figure 3.5 shows the nucleotide multiple alignment of all three TCs with the translated protein sequence (ISCW002300-PA) of Histone H3 transcript derived from *I. scapularis* (ISCW002300-RA). While TC1346 aligns with TC1345 in a colinear fashion over the entire length of the histone coding region TC1347 aligns with TC1345 in two segments separated by 664 bp. The two aligned segments have 97% and 99% sequence identity, respectively with TC1345. TC1347 is assembled from six ESTs of which five span the second segment that aligns with TC1345 while one spans both the segments, suggesting that the gap observed in the alignment of TC1347 with TC1345 is not likely to be a result of an assembly artefact.

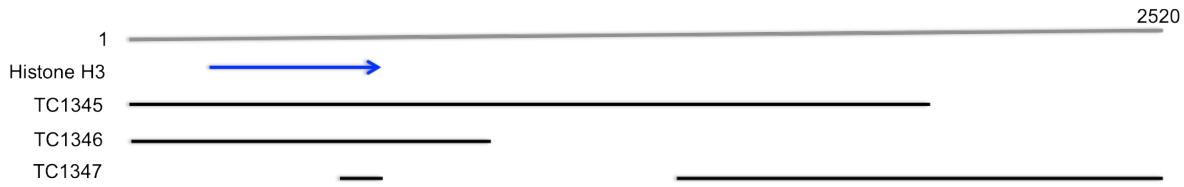


Figure 3.5 Diagrammatic representation of how sequences in cluster TC1345-1347 align relative to each other. Black lines represent aligned nucleotides with breaks indicating alignment gaps. TC1347 aligns to TC1345 and TC1346 in two segments. Blue arrow represents the 5' to 3' orientation and the region homologous to Histone H3 protein (ISCW002300-PA) of I. scapularis. The figure is to scale.

© University of Pretoria

Homologues of Histone H3 are present in all other tick EST databases however the variant TC1347 is unique to *R. appendiculatus*. A fourth sequence putatively coding for a variant of Histone H3 is present in RaGI. TC949 encodes a 70% partial ORF closest in sequence identity (76% nucleotide identity) to *Drosophila teissieri* Histone H3 (GB Accession: AB019404.1). It has 73% nucleotide sequence similarity with TC1345 and TC1346 over ~400 bp. The *I. scapularis* genome assembly has four paralogous copies of the Histone H3 gene (ISCW001838, ISCW002300, ISCW002893 and ISCW003178). The ISCW002300 gene is the closest in sequence identity to the TC1345-7 cluster. Three of the *I. scapularis* Histone H3 genes have two exons and one intron that is longer than 1 Kb. The ISCW002300 gene has two exons (282 bp, 129 bp) separated by a 2382 bp intron, 50 bp 5' and 3' UTRs, and flanked by 69 bp upstream and 45 bp downstream intergenic regions. The alignment depicted in Figure 3.5 indicates that Histone H3 genes in *R. appendiculatus* have 3' UTRs of varied lengths. It has been shown in *Drosophila* (Akhmanova et al., 1995) and the domestic chicken (*Gallus domesticus*) (Dodgson et al., 1987) that there is variation in the 3' UTRs of Histone H3.3 genes. The three RA Histone-containing TCs possess long 3' UTRs - in TC1345 its length is 1386 bp, in TC1346 288 bp and in TC1347 there are 1271 bp in the 3' UTR region. UTRscan (Pesole, 2000) identifies several putative regulatory elements located in the 3' UTR of TC1345 as well as in its *R. microplus* homolog, TC19312, namely Cytoplasmic Polyadenylation Element (CPE), Internal Ribosome Entry Site (IRES), Sex-lethal binding site (SXL_BS) and Brd-Box (only in TC19312). Functional data shows that SXL_BS and Brd-Box are involved in regulation of transcript expression in Drosophila (Beckmann et al., 2005). Our result suggests that the 3' UTRs of ixodid ticks may play a role in regulation of some translation processes.

3.4.1.4 TC9-11 cluster

An open reading frame of 1034 aa contained within TC10 has an average protein sequence identity of 91% with the alpha subunit of sodium-potassium ATPase from a range of arthropod species, including *I. scapularis* (black legged tick), *Tribolium castaneum* (red flour beetle), *Callinectes sapidus* (blue crab), *Apis mellifera* (honey bee), several *Drosophila* species and *Pediculus humanus corporis* (human body louse)

© University of Pretoria

amongst others. This gene is also conserved in many vertebrates including humans, apes, cattle, canines, rodents, chicken and zebrafish.

Sodium-potassium ATPase is an integral membrane protein essential for the function of the sodium ion pump that maintains ionic gradient across the cell membrane. It belongs to the P-type ATPase group to which ATPases such as proton pump, proton-potassium pump and calcium pump belong. The sodium pump is composed of two subunits - α (~1000 aa, ~110 kDa) and β (~300 aa, ~55 kDa) - encoded by distinct genes. The α subunit contains the functional sites and is responsible for the enzyme's catalytic activity. Four isoforms of the catalytic α subunit and three for the β subunit have been identified in mammals. In the crab, *Callinectes sapidus,* the protein folds into eight transmembrane helices (Towle et al., 2001). While most subfamilies of Na+/K+ ATPase have ten transmembrane helices, variation in this number is frequently observed in some subfamilies. The gene putatively encoding the α subunit of the Na+/K+ ATPase of *I. scapularis* (ISCW002538) contains 20 exons. Topology prediction using TMHMM (Sonnhammer et al., 1998) predicts six transmembrane helices for ISCW002538 and eight for TC10, implying that the two tick proteins differ in the specific details of how they associate with tick membranes.

TC10 aligns with ISCW002538, TC9 and TC11 as shown in Figure 3.6. TC10 overlaps with TC11 over 1,027 bp with 98.5% nucleotide identity downstream of the Na+/K+ ATPase coding sequence. TC9 does not align with TC10 but it aligns with TC11 ~800 bp from its 3' end with 95% nucleotide identity. TC9 contains a ~700 bp insertion when compared with TC11, the region flanking it being ~98% identical. The region ~5 kb downstream of the Na+/K+ ATPase coding sequence covered by TC11 and TC9 presumably contains the 3' UTR of the gene.

The UTRs of mRNAs of Na+/K+ ATPase subunits have been shown to regulate translation in mammals and the length of the 3' UTR is thought to be inversely proportional to the translational efficiency (Clifford and Kaplan, 2009; Shao and Ismail-Beigi, 2004). Part of the ~2 kb region between ISCW002538 and ISCW002540, the locus

© University of Pretoria

downstream in the draft assembly of *I. scapularis* genome could also be the 3' UTR of the Na+/K+ ATPase gene. Thus there may also be post-translational regulation of this transporter gene family in the prostriate tick lineage.

Figure 3.6 Diagrammatic representation of how sequences in cluster TC9-11 align relative to each other and to the I. scapularis homologue ISCW002538_RA. Black lines represent aligned nucleotides with thin black lines indicating alignment gaps. The figure is to scale.

A partial homologue of TC10 is present in AvGI (BM290953) but absent in BmiGI indicating that despite normalisation prior to library construction the EST dataset for this species is missing essential protein coding genes that are likely to be present and again, highlighting the technical biases inherent in different methods of generating EST datasets. BmiGI TC15685 has > 80% nucleotide sequence similarity with TC9 over ~1,200 bp, suggesting conservation of non-coding regions, but it has no sequence similarity with TC10.

TC11 is 3,985 bp in length and is assembled from a large total of 70 ESTs. Six of these ESTs have a visible attached polyT or a polyA tract and according to the method of library construction the remaining transcripts within the cluster should also be polyadenylated. The sequence contains numerous A/T rich stretches. TC11 contains ten predicted ORFs longer than 100 aa, the longest of them being 228 aa. In BmiGI TC11 has sequence identity with TC2159 and TC179, both of which are unannotated and have six and one ORFs longer than 100 aa, respectively. There are three singletons homologous to TC11 in AvGI, one EST match in the *A. americanum* EST dataset, and two EST matches in the *I. scapularis* salivary gland EST dataset. Although it does not appear to encode a classical Na+/K+ ion transporter the sequence similarity at the 5' end might suggest that one of its roles is in regulation of this pathway.

© University of Pretoria

3.4.1.5 TC3-5 cluster

Three TCs, TC3, 4 and 5 comprise this cluster. They have a complex arrangement and may represent ncRNA. They are discussed further in Chapter 4, Section 4.2.2.

3.4.1.6 TC1286-93 cluster

This group consists of six TCs and one singleton and is discussed in detail in Chapter 4, Section 4.2.3.

3.4.2 Novel Putative Immune Regulatory Molecules in *R. appendiculatus*

3.4.2.1 TC1324-6 cluster

This group consists of three TCs. Members of this group are unannotated and have no sequence homology to known proteins. TC1324, 3,656 bp in length, contains a long ORF comprising 1,133 aa. It has a polyadenylation site and one EST has a visible attached polyT tract. The ORF begins with an in-frame methionine and has an average G+C content of 49.3%. The predicted protein sequence is glycine-rich. A signal peptide was predicted for the ORF; using the HMM method the probability score was 0.999 with a cleavage site probability of 0.583 between residues 20 and 21.

Within RaGI, TC1324 is homologous to TC1325 and TC1326. Figure 3.7 shows an outline of the alignment between the three RaGI homologues. TC1324 has 94% and 87% identity with nucleotide sequences of TC1325 and TC1326, respectively.

Figure 3.7 Diagrammatic representation of how sequences in cluster TC1324-6 align relative to each other. Grey bars represent aligned nucleotides, horizontal lines within them indicating gaps. Numbering above the diagram marks the nucleotide position of the sequence. Figure is to scale.

© University of Pretoria

BmiGI TC1135 and singleton CK175673 have sequence identities of 84% and 83%, respectively with TC1324 over 1,109 bp and 664 bp, respectively, at the 3' end of TC1324. Similar sequences to the TC were not identified in AvGI. A BlastX search of TC1324 against GenBank NR database resulted in no significant matches. The ORF did not have similarities to any known conserved domains based on a search of the sequence against the Genbank Conserved Domains Database (CDD) either. This could therefore be a protein or protein family specific to the genus *Rhipicephalus*.

A protein of this length is likely to contain multiple domains. To investigate this putative domains in the ORF were predicted using the DomPred pipeline, which utilizes PSI-BLAST, PsiPred and DomSSEA sequentially. Although the top four results from DomPred all predicted four domains the domain boundaries as well as the SCOP codes associated with the hits differed in three of the four domain matches observed using this algorithm. Owing to the lack of consistency in the results generated using the DomPred pipeline a second algorithm, HHPred, which identifies conserved structural modules for the ORF was used. Three putative domains were identified using this method.

The domain with the highest probability for the template to be a true positive (Prob: 65.4) was located between residues 609 and 643 of TC1324 between residues 311-345 of the crystal structure of the C4d fragment of complement factor C4 of humans (PDB: 1hzf_A; van den Elsen et al., 2002). This *R. appendiculatus* domain also had hits to several other structures for complement fragments from mammals.

A tertiary structure model for the ORF was generated using the I-Tasser web service. The best model was predicted with a confidence score of -0.68 and a TM-score of 0.63+-0.14. A search of the predicted 3D model against the PDB database was performed using the Dali server. Matches with a Z-score of above 10 aligned within the C-terminal region of the 3D model of the protein from residues 1,017 to 1,131. The top three hits, which had Z-scores between 11.8-11.6, were to structures corresponding to Complement Factor H (CFH) in humans. CFH is involved in regulation of the alternative pathway of

© University of Pretoria

complement i.e. it protects self cells from complement activation. Several microbes have been shown to express Factor H binding protein (FHbp), which recruits host CFH to its surface thereby evading the alternative pathway by inhibiting complement-mediated lysis (Haapasalo et al., 2012; Schneider et al., 2009). Factor H interacts with the activated C3, C3b, to inhibit the alternative complement pathway. Other hits with Z-scores above 10 were to Complement Factor C3, suggesting possible mimicry of both complement regulatory proteins and complement itself by *R. appendiculatus*.

Complement inhibition in metastriata has not been reported previously. However, saliva of several prostriate (Ixodes) ticks has been shown to inhibit the alternative pathway (Lawrie et al., 1999; Ribeiro, 1987; Wikel and Allen, 1978). Additionally, antigenically variable anti-complement molecules, Ixodes AntiComplement (IxAC), which inhibit C3 convertase by binding to preprodin, have been found to be specific to the *Ixodes* genus (Couvreur et al., 2008; Valenzuela et al., 2000).

The consistency in matches to complement components obtained with putative structural homologues generated using different prediction algorithms strongly suggests that TC1324 could be expressing a complement pathway inhibitory molecule that is specific to the *Rhipicephalus* genus and whose inhibition mechanism may involve regulating the binding of Factor H with C3. This protein has a potential to be a novel acaricide target against Rhipicephalus species as it appears to be specific to the genus and acts at a very early stage in the alternate pathway.

3.4.2.2 Identification of a Variabilin homologue in *Rhipicephalus appendiculatus*

In another study undertaken at ILRI, cDNA prepared from RNA that was extracted from *T. parva*-infected and uninfected *R. appendiculatus* salivary glands was used as template for Suppression Subtractive Hybridization (SSH). SSH is a technique that enables identification of differentially expressed genes. Several differentially expressed transcripts were identified and the differential expression was validated using Real

© University of Pretoria

Time PCR. One of the validated genes, designated clone E4 was approximately 300% up-regulated in infected RA salivary glands relative to uninfected. It was amongst the few differentially expressed genes in this study that didn't have primary sequence identity to any known proteins in the GenBank NR database and hence more sophisticated *in silico* methods were employed to elucidate its function.

The longest ORF within the E4 transcript that was 268 bp in length contained 69 residues. It has a predicted signal peptide based on the Neural Network and Hidden Markov Model (HMM) methods of SignalP. Sequence homology based methods gave no significant matches to genes present in GenBank. In order to attempt to assign a function to this gene the fold recognition method GenThreader was employed. GenThreader uses a sequence-structure alignment generated using a sequence profile method. It evaluates the alignments using pairwise and solvation potentials, then feeds the scores into a Neural Network for evaluation and generation of an overall score.

The best matches for the 69-residue ORF in E4 using GenThreader are to disintegrins - decorsin from leech (PDB: 1DECA; Krezel et al., 1994) and trimestatin, found in snake venom (PDB: 1J2L; Fujii et al., 2003). Although the assigned scores were not significant (Probability of false positive > 0.1; SVM score between 23-24) conservation of certain domains, as is evident in the sequence alignment (Figure 3.8), along with the nature of the functions of the protein matches indicated that further investigation would be worthwhile.

```
                                  10            20         30
SS       ---------------CCCCCCCCC----CCCCCEEC----CCCEECCCCCEE-CCCCCCC--CEEC-----
1decA0   ---------------APRLPQCQG----DDQEKCLC----NKDECPPGQCR-FPRGDAD--PYCE-----
                       |  |   ||         | |     |  || | |   ||||    | |
E4       MKAAYLLALTALLILATSL-GIQGSHISEGGNPCDCVSQEAKEACPDGHCACWPRGDTPEGPTCIPQRSK
                10         20        30        40        50        60
```

Figure 3.8 Top GenThreader match (PDB: 1decA0), in alignment format, of E4 ORF.

Decorsin (39 aa) and trimestatin (70 aa) are antagonists to Glycoprotein (GP) IIb/IIIa that serves as the final common pathway to platelet aggregation. GPIIbIIIa, also know as integrin αIIbβ3, is an integrin located on the surface of blood platelets. There are about 50,000 GPIIbIIIa on a single platelet. GPIIbIIIa serves as a receptor for at least four

© University of Pretoria

plasma protein ligands – fibrinogen, fibronectin, vitronectin and von Willibrand factor. These ligands contain the tripeptide integrin recognition motif, RGD (Arg-Gly-Asp), which mediates cell attachment. RGD motif of fibrinogen binds to active GPIIbIIIa of one platelet, which then binds to an active GPIIbIIIa of another platelet forming a clot. Disintegrins that are antagonists of GPIIb/IIIa have been shown to inhibit platelet aggregation. Drugs consisting disintegrins are used for stroke prevention. E4 ORF1 and decorsin share amino acid identity of approximately 43%.

Other GPIIa/IIIb antagonists containing the RGD motif have been isolated from leeches (ornatin), snakes (disintegrins, kistrin), soft ticks (Savignygrin – Mans et al., 2002) and the hard tick *Dermacentor variabilis* (variabilin). Variabilin (Wang et al., 1996) is a 47 aa protein containing the RGD motif. It was shown to have antiplatelet activity. Variabilin has a global protein identity of 36.6% to E4 ORF1. There is a putative homologue in *Rhipicephalus sanguineus* (ACX53898.1), which has 75.7% global amino acid identity with E4 ORF1 and 30.6% with Variabilin. Figure 3.9 shows a multiple sequence alignment generated using Muscle of the three disintegrin proteins isolated from hard ticks *D. variabilis, R. sanguineus* and *R. appendiculatus*. The secondary structure of the E4 protein, predicted using PSIPRED v3.0 is also shown on the second row of the alignment, along with the confidence values of predictions for each residue (first row). The predicted signal peptide in E4 is underlined and the cleavage sites are in bold. The RGD motif (boxed) is conserved in all three proteins however, a single nucleotide variation in the RGD motif of *R. sanguineus*, possibly resulting from erroneous base calling, brings about a non-synonymous change of R to K (although both of these are basic amino acids, therefore functionally similar). The KGD, along with RYD and OrnGD motifs have been shown to be recognized by ligands (Mans et al., 2002), whereas a change in this motif from RGD to RAD has been shown to block the binding activity in human K562 cells (Barry et al., 2000). The RGD motif is located in a surface loop of the protein. Loops tend to be the most variable parts of a protein. The conservation of RGD and eight residues surrounding it, all of which are predicted, with high confidence, to be located in the loop suggests these residues could have important roles in the protein function, most likely in interaction with other molecules.

© University of Pretoria

```
MUSCLE (3.6) multiple sequence alignment

Confidence      92788999999999997633345-313789--998767786788649999300046999 9
Prediction      CHHHHHHHHHHHHHHHHHHCCCC-CCCCCC--CCCCCCCCHHHHHHCCCCCEECCCCCCC 9
Variabilin      ------------------------NTFSDENPGFPCDCTSADALRAC-GIQCACWPRGDT 3
ACX53898.1      MKAAYLLTLTALLILATSMGVQGWSHISEG--KNPCDCESPEAQQACPHGHCACWPKGDT 5
E4_orf1         MKAAYLLALTALLILATSLGIQG-SHISEG--GNPCDCVSQEAKEACPDGHCACWPRGDT 5
                                       . :*:     **** * :*  **   :*****.*** 

Confidence      999822446889 9
Prediction      CCCCCCCCCCCC 9
Variabilin      PGGGRRIIDGQQ 3
ACX53898.1      PGGPKCIPK--- 3
E4_orf1         PEGPTCIPQRSK 3
                * *   * .
```

Figure 3.9 Multiple alignment of tick disintegrin sequences with closest protein identity to E4 - D. variabilis (Variabilin), R. appendiculatus (E4) and R. sanguineus (ACX53898.1). RGD motif is highlighted in grey. Signal peptide is underlined.

The RGD epitope of fibrinogen inhibitors decorsin and ornatin from leeches are flanked immediately by Cys residues that form a disulfide bridge causing the RGD epitope to protrude into a loop for presentation. As is observed in Decorsin and some snake disintegrins such as Kistrin (Adler et al., 1991; Wang et al., 1996) the variabilin RGD motif is flanked by Proline residues suggesting that RGD presentation in these molecules occurs in a different way compared to the Cys-flanked RGD molecules (Wang et al., 1996) A ligand that was well-projected from the surface of the protein was shown to have a stronger bind to the receptor compared to one that was slightly hidden (Lee et al., 1993). Positioning of the RGD motif in a projecting loop of decorsin suggests that a mechanism other than formation of a disulphide bridge to enable protrusion of the RGD motif may be at play. This is arrangement of the RGD motif is also seen in the model of E4/F12 threaded onto decorsin suggesting that E4/F12 is likely to have high affinity to integrins. Figure 3.10 shows the protein structure model of E4 in a static state. The protruding loop is highlighted in green. The side chains of the R, G and D residues are marked in blue, white and red, respectively.

© University of Pretoria

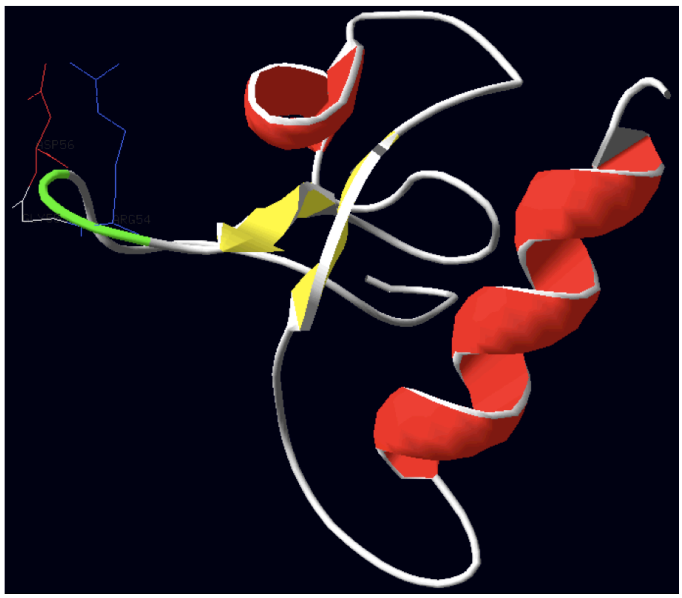


Figure 3.10 Protein structure model of E4 (static view). The protruding loop is highlighted in green. The side chains of the R, G and D residues are marked in blue, white and red, respectively.

The absence of E4 in RaGI is not surprising because the RNA samples used for SSH were not size-selected, unlike those used for generation of RaGI. It is important to note that using the frequently used cut-off length of 100 aa to identify ORFs employed by automated annotation algorithms may miss short, functionally important proteins such as E4. This therefore may need to be modified for tick peptides.

Proteins that modulate the host immune system have been previously isolated in ticks. E4 represents an additional protein that has the potential to protect the tick from the arsenal of immune responses elicited by the host upon a tick bite.

3.5 Glycine rich proteins

Glycine-rich proteins make up a multi-functional heterogenous group that are highly prevalent in salivary gland EST datasets. In ticks, these proteins are abundantly present in the tick cement cone that is thought to function as attachment glue facilitating the long periods of attachment to the mammalian hosts. They were also found to be up-

© University of Pretoria

regulated by blood-feeding in *I. ricinus* (Chmelař et al., 2008). Tick glycine-rich proteins isolated from the cement cone have been shown to have immunogenic properties. They may inhibit platelet aggregation by interacting with platelet collagen receptors (Ribeiro et al., 2006). RIM36 already described above is immunogenic in the mammalian host and invokes strong antibody responses in cattle infested with *R. appendiculatus* ticks (Bishop et al., 2002). Although recombinant forms did not induce protection against tick infestation in cattle (Bishop and Musoke, personal communication).

3.5.1 Glycine-rich proteins in tick Gene Indices

A total of 210 glycine rich sequences, with glycine amino acid content > 23% were identified within the tick EST databases. RaGI contains 92 such sequences, while there are 42 sequences in BmiGI, 71 in IsGI, but only 5 in AvGI, perhaps consistent with the lower number of transcripts present in this EST database. Of the 92 RaGI glycine-rich sequences identified tentative annotations were assigned to 42 sequences. 47% of the annotated RaGI glycine-rich protein (GRP) sequences are annotated as glycoproteins, consistent with the heavy glycosylation of this category of sequences. Other annotations include proteins similar to keratin, silk proteins and cell-wall proteins. Although as already mentioned, these tentative annotations should be interpreted with caution due to the skewed amino acid composition of the predicted proteins.

Domain searches against the Interpro database resulted in significant matches to five domains – LIM, Eggshell, RNA Recognition Motif (RRM), Zinc finger RanBP-type and Chitin-binding type R & R domains - in 33 of glycine rich sequences.

3.5.1.1 LIM domain

The LIM domain is named after the proteins in which they were first identified – Lin11(L), Isl-1(I) and Mec-3(M). Proteins containing the LIM domain have been shown to be involved in organisation of the cytoskeleton. The sequences encoding LIM domains are highly divergent except for a few conserved residues that are present within zinc-finger domains.

© University of Pretoria

This domain is present in four tick EST sequences; three from IsGI and one from BmiGI. The cysteine residues are conserved in all four sequences. No sequences in RaGI contained a LIM domain. This might indicate that LIM mediated cytoskeleton regulation is less prevalent in salivary glands relative to other tick tissues.

3.5.1.2 Eggshell-domain

The Eggshell domain first identified in *Schistosome mansoni*, is rich in Gly and Tyr residues. Seventeen sequences containing this domain were identified in the tick EST databases; ten from RaGI, four from BmiGI, two from AvGI and one from IsGI. The presence of this domain in RaGI is interesting, since the function in the salivary gland is presumably not related to eggshell formation.

3.5.1.3 RNA Recognition Motif (RRM)

The RRM motif is one of the most abundant protein domains in eukaryotes. It is found in proteins contained within ribonucleoprotein complexes involved in post-transcriptional gene expression processes. The domain contains two highly conserved regions – Ribonucleoprotein (RNP)-2 and RNP-1 – that are thought to be responsible for binding of RNA molecules (Maris et al., 2005). In mammals the motif is found in heterogeneous nuclear ribonucleoproteins (hnRNPs), small nuclear ribonucleoproteins (snRNPs), pre-RNA and mRNA-associated proteins, plus RNA-processing proteins. In the tick databases six sequences contained the RRM domain. Four of the sequences are very closely related to each other - CD794403, TC1768 from RaGI, TC21505 from BmiGI and TC55577 from IsGI. A multiple alignment of the four sequences is show in Figure 3.11. The RRM domain and a zinc finger domain that are present in all of these TCs are indicated in the figure.

The protein sequence is remarkably conserved in all three species of ticks. In contrast, the nucleotide sequence is very diverse, which is consistent with ancient divergence of metastriate and prostriate lineages, but strong evolutionary pressure for functional conservation of the protein. The conservation of this domain suggests that the similarity

© University of Pretoria

between ticks and vertebrates can sometimes extend beyond genome size and overall organization to functional aspects of gene expression control.

Figure 3.11 Multiple sequence alignment of glycine-rich sequences from R. appendiculatus (RaGI_CD794403, RaGI_TC1768), R. microplus (BmiGI_TC21505) and I. scapularis (IsGI_TC55577) that contain the RNA Recognition Motif (RRM) (red box) and zinc-finger domain (blue box).

3.5.1.4 Zinc finger Ran-Binding Protein-type

This domain is usually found in proteins that are involved in nuclear transport. Four glycine rich protein sequences (two from RaGI, one from BmiGI and one from IsGI) were found to contain this domain.

© University of Pretoria

3.5.1.5 Chitin-binding domain

This domain occurs in soft/flexible cuticle proteins of insects and arachnids. It was present within two transcripts in RaGI, suggesting wide evolutionary conservation in the organization of the arthropod cuticle.

3.5.2 Clustering based on amino acid residues and presence of conserved amino acid motifs

The 210 glycine-rich proteins within RaGI were clustered into groups of similar sequences based on conserved repeated amino acid motifs. Conserved glycine rich repeat motifs were predicted using MEME (Bailey and Elkan, 1994), which uses statistical modelling techniques applied to a group of related sequences to assemble the clusters. Seventeen conserved motifs were identified in the clusters. The GGY motif was the most abundant, and was present in 47 sequences, with repeat copy numbers ranging from 5 to 15. Table 3.4 lists all the motifs, the number of sequences in which that motif is present and the number of copies.

Table 3.4 Glycine-rich motifs identified in RaGI.

Motif	No. Sequences	Range of no. copies
GGY	47	5-15
YGSSGL[GS]GLG[YF]G[SG][YS]G	7	1-13
YG[GS]YG[SG]GLG[GS]YG	12	1-9
[RS]GLGSLGFGSGS	10	2
[GN][TL][SP][GQ][VE][GR][RA][GS][VM][TK][GV]FV[LG][IP]	19	1-3
ALG[GA]LP[GV][GA]A[VA]GVLPS	19	1
ASS[VA][GSA][VL]R[PL]GSA[GV]RG[AV]	19	1
[PS][ST][AV]GVG[GS][LF][AS]GGS[FL]GP	18	1
GYGPYGG[VG]FGNAGYW	8	1
[FAIV]GVPLFGGY	10	1
GV[SR]ATG[SG]	8	1
GSAS[GA]S[LP]GAVG[RS][VG]G	5	1

© University of Pretoria

3.6 Conclusion

Manual curation has brought about an improvement in the understanding of the functional categories represented within the *R. appendiculatus* salivary gland transcripts. Novel peptides potentially having immune-modulatory functions have been brought to light, such as the proteins containing Kunitz domains, TC1324 and E4.

Obtaining novel transcripts from a second library made from the uninfected salivary gland material indicates that many genes transcribed by *R. appendiculatus* remain to be discovered. Sequencing of additional clones may also alter the initial observation made by Nene et al. (2004) that infected and uninfected salivary glands are not differentially expressed. The latter point is confuted by the finding that the expression of E4 was 300% up-regulated in infected salivary glands.

Size selection of expressed transcripts also prevents identification of transcripts coding for short proteins such as defensins, which tend to be 34-46 aa long (Rudenko et al., 2007), and variabilin, which is 69 aa in length. Furthermore, deep sequencing would resolve sequencing errors such as those seen in the *R. appendiculatus* homologues of serpins.

© University of Pretoria

Chapter 4. Non-coding RNA in transcribed sequences

4.1 Background

Recent studies have shown that large portions of non-coding regions in genomes are transcribed (Mattick, 2003; Mattick, 2005). Most of these studies have been carried out in humans and other mammals. Very little data is available for arthropods, although many arthropod genomes, particularly those of ticks (Sunter et al., 2008) are similar in non-coding content, to those of mammals. In humans, non-coding RNA (ncRNA) accounts for approximately 98% of all genomic transcripts (Mattick, 2003). Fifty six percent of the cytosolic transcriptome (mature RNA) lacks matches with annotated genes whereas 80% of the nuclear transcriptome derives from the unannotated section of the genome (Kapranov et al., 2007a). Long considered as non-functional, ncRNAs are now being found to play regulatory roles in genes. The best-characterised ncRNAs are those that are present in relatively large quantities and associated with known or putative functions. These include transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), micro RNAs (miRNA), small interfering RNAs (siRNA), small nucleolar RNAs (snoRNA) and small nuclear RNAs (snRNA). However, there is also an abundance of longer ncRNAs expressed at lower levels (Mattick and Makunin, 2006).

ncRNAs could be responsible for some of the phenotypic differences observed between species. As only 0.3% of sequence variations observed between individuals occurs in the protein-coding regions in humans this class of RNA is suspected to influence phenotypic divergence between individuals within a species (Claverie, 2005; Mattick, 2001; Mattick, 2004).

ncRNAs in eukaryotes are thought to have been derived from introns and other non-protein-coding sequences in the genome (Mattick 2001). In humans, $^{2}/_{3}$ of all the

© University of Pretoria

transcripts are non-protein coding and approximately 95% of the pre-mRNA transcripts are intronic ncRNAs. In *C. elegans* almost 50% of the ncRNA loci are intronic (He et al., 2006). Most ncRNAs that have been identified have not yet been characterised in detail and have no function identified for them. It has been suggested that some non-protein coding RNAs could function as trans-acting regulatory signals (Kohtz et al., 2006; Mattick, 2001). In prokaryotes, although they are much less frequent than in mammals a number of trans-acting small ncRNAs have been described that regulate mRNA translation and stability (Mattick and Makunin, 2006).

In eukaryotes, one suggested function of ncRNA is regulation of transcription or translation of proteins by RNA (ribo-regulation) (Galindo et al., 2007). The regulatory functions described to date are mostly implemented by ncRNA encoding base pairs complementary with sequences of other RNA/DNA molecules enabling the formation of RNA:RNA or RNA:DNA complexes that are recognised and subsequently trigger the activities of other complexes. ncRNAs appear to be multifunctional and also play a role in chromosome maintenance and segregation and some are involved in cell biological processes as well as in RNA editing, translational inhibition, mRNA destruction and stress responses. They may also act as scaffolding for the assembly of macromolecular complexes (Mattick and Makunin, 2006). Most non-coding transcripts are cell-type specific, and developmentally regulated. One example being snoRNAs whose failure to function properly has been implicated in many diseases. RNA signals are known to be fundamental in underpinning certain gene regulatory and eipigenetic phenomena including RNA interference (RNAi), DNA methylation and transgene silencing. RNAi is well known for its use in defense against viruses and transposons, but other processes are also crucial in cell development, for example RNA regulation of chromatin architecture and modulation of transcription (Mattick, 2001).

In the mammalian genome, an estimated 22% of human transcription clusters analysed, form sense-antisense pairs. Most of these antisense transcripts represent ncRNA. Analysis of the mouse transcriptome indicates that a high proportion amounting to 72%, of the transcriptional units can overlap with transcripts from the opposite strand

© University of Pretoria

(Mattick and Makunin, 2006). A number of antisense transcripts appear to play regulatory roles through control of expression of the corresponding sense gene. The presence of ncRNA has also been demonstrated in single celled organisms. Tiling arrays have identified a number of antisense and intergenic transcripts in *E. coli* (Claverie, 2005). In yeast (*Saccharomyces cerevisiae*), a large number of ncRNA sequences also show partial overlap with ORFs encoded on the opposite strand, and these are usually conserved in sequence (Havilio et al., 2005). Antisense transcripts have also been reported in ixodid ticks (Chmelar et al., 2008).

In mammals, some non-coding sequences appear to be more highly conserved than protein coding sequences (Mattick and Makunin, 2006). Many miRNAs that control aspects of plant and animal development through sequence-specific interactions with other RNAs are highly conserved. miRNAs appear to have multiple targets thereby promoting evolutionary conservation. However this is not always the case, for example, *lys-6* in *C. elegans*, which plays a key role in determining left/right neuronal asymmetry, has limited evolutionary conservation. miRNAs *X*ist and Air in mammals are also poorly conserved. In mammals, the upstream regions of ncRNA transcripts show many of the features normally associated with promoters and may be more highly conserved than the promoters of protein-coding genes (Mattick and Makunin, 2006).

A majority of the non-coding transcripts in humans occur as RNAs longer than 200 bp, and are referred to as lRNA (Kapranov et al., 2007b). Some examples include Xist (17kbp), Evf (2.7kbp), Air (3.7kbp). A large number of ncRNAs are also < 200 bp in length, and are referred to as short RNAs (sRNA). sRNAs were found to be abundantly located at the termini and the coding regions of genes in humans (Kapranov et al., 2007a, Kapranov et al., 2007b). They were also found to be expressed at similar levels to the coding genes that they overlapped. Some sRNAs were found to overlap with lRNAs, suggesting that lRNAs could be precursors for sRNAs.

Some ncRNAs have been found to have multiple polyadenylation sites (polyA) sites and are alternatively spliced. NRON, which is a repressor of the transcription factor NFAT in mouse, exists as a series of alternatively spliced transcripts ranging from 0.8 to 3.7 Kb

© University of Pretoria

(Mattick and Makunin, 2006). This suggests that they could have been derived from protein-coding sequences that have lost this function during evolution.

Distinguishing ncRNA from mRNA remains a challenge. Discriminating by ORF length alone could result in false classifications. Short ORF-containing RNA (ORFs shorter than 100 aa) that encode proteins have been observed in multiple species, with the mammalian proteome estimated to contain 3,700 proteins below this size. Combining ORF length with similarities to known functional proteins is likely to increase ncRNA prediction by elimination of functional mRNA. However, this is influenced by the number of known protein coding genes in the public databases and the quality of their annotations.

To add to the challenge of classifying RNA as non-coding there is increased evidence of transcripts that can be both functional at the RNA level as well encode proteins (Dinger et al., 2008). These bi-functional RNAs have been identified in humans (Steroid Receptor Activator (SRA)), *Xenopus* (*VegT*RNA), *Drosophila* (*Oskar* RNA) and plants (*enod*40) among other taxa. A bi-functional RNA could either be targeting its protein-coding counterpart or have a different target site altogether, enabling regulation of either the same pathway as the coding section or modulating an entirely different network of genes. In the case of SRA, the functional protein SRAP, acts antagonistically to the non-coding component of the transcript (Dinger et al., 2008).

It appears that the occurrence of ncRNA is widespread in higher eukaryotes, including mammals and plants and is functional in a variety of different processes. It also occurs in unicellular eukaryotes and to a lesser extent, prokaryotes. ncRNA has to date been less well studied in arthropods, particularly ticks.

I analysed RaGI for the presence of potential non-coding RNA transcripts. I observed a lack of significant sequence similarity to sequences in the databases and absence of ORFs longer than 100 amino acids in *R. appendiculatus* assembled polyadenylated

© University of Pretoria

transcripts. Given its widespread abundance it therefore seems probable that a significant proportion of these sequences are likely to be ncRNA.

4.2 Evidence of non-coding RNA in RaGI

Of the 957 unannotated TCs approximately 61% contain ORFs longer than 100 aa. The average length of these TCs is 1,082 bp ranging from 3,985 to 343 bp. The longest, TC11, is 3,985 bp and is assembled from 70 ESTs. In RaGI TC11 is homologous to TC9 and TC10 (discussed in Chapter 3).

Amongst TCs containing no predicted ORFs of more than 100 amino acids in length TC1286 had the highest sequence redundancy comprising 38 ESTs. Thirty-seven percent of the TCs within this category were assembled from only 2 ESTs. TC4 is the longest having 2,495 bp and TC948 is the shortest, made up of 244 bp. The average G+C composition of these 362 TCs having short ORFs is 41.53%, varying from 54.5% to 33.7%.

The non-coding probability of RaGI sequences was predicted using PORTRAIT. Of the 7,340 total assembled RaGI TC and singleton sequences 2,394 were predicted to have non-coding regions (more than 77% of these are singletons). Among the multi-copy transcripts having high predicted non-coding probability, 130 TCs were tentatively annotated and 405 TCs were unannotated. The annotations assigned to the 130 TCs were assessed to find that they included ribosomal proteins, RNA polymerase as well as proteins such as tropomyosin and phosphoglucose isomerase. It was surprising to find that transcripts coding for known protein-coding genes were being predicted as being non-coding. Further investigation into these supposedly false positives revealed that the non-coding properties of these transcripts were located in the sequence flanking the protein-coding region of the transcript. For example, in *R. appendiculatus*, tropomyosin is encoded by ~300 bp at the 5' end of TC1370, the TC being 1,043 bp in length. In mammals, *mir-21,* a 23 bp miRNA that regulates tropomyosin 1 (*tpm1*), was found to be

© University of Pretoria

located in the 3' UTR of tropomyosin gene (Zhu et al., 2007). It is possible that the ~750 bp downstream of the TC encoding tropomyosin in RaGI contains the 3' UTR.

Interestingly, approximately 50% of the predicted ORFs of unannotated TCs clustered at the termini indicating that remaining sequence lacking long ORFs is likely to be non-protein coding, probably a UTR. In some cases the regions of sequence lacking long ORFs are longer than the protein-coding domains. Similar ORF distributions in ESTs were observed in a bumblebee *Bombus terrestris* EST database (Sadd et al., 2010).

118 of the predicted noncoding sequences are unique to RaGI, i.e. they show no significant match to sequences in GenBank. The majority of these TCs have low transcription levels – 114 are assembled from five or less ESTs, one with seven ESTs, two with 10 ESTs and one with 49 ESTs. 38 of these TCs are longer than 1 kb, and only one TC is < 300 bp in length. The majority of the putative non-coding TCs (106) had a G+C content below the average 50% observed in protein-coding genes of *R. appendiculatus*.

Based on the observation that the average length of proteins in the Swissprot database is 100 aa, a cut-off that can be used for initial classification of transcripts, protein encoding sequences should contain at least one 300 bp ORF, although in some cases this will exclude genes encoding short peptides. Dinger et al., 2008 demonstrate using mathematical models, that the mean ORF length expected for a transcript 1,000 bp long is approximately 170 aa. Taking these factors into account I analysed 957 unannotated TCs for the absence of ORFs exceeding a certain length in three different categories of ORF lengths.

Three lists of TCs, having no ORFs longer than 120 aa, 100 aa and 50 aa, respectively, were compiled. Five hundred and ninety three TCs were observed to contain no ORFs longer than 120 aa (Table 4.1), 390 TCs had no ORFs longer than 100 aa (Table 4.2) and 3 TCs had no ORFs longer than 50 aa (Table 4.3). The tables list TCs with the highest levels of expression, using EST redundancy in a TC as an indicator. It was assumed that

© University of Pretoria

the higher the EST counts within a TC, the higher the expression level of the gene encoded by the TC, since the cDNA library used to generate RaGI was not normalised.

Table 4.1 Summary of properties of 10 of the 593 tentative consensus sequences (TCs), arranged in descending order of number of ESTs within the TC, containing no Open Reading Frames (ORFs) more than 120 amino acids (aa) in length. Location of the longest ORF within the TC sequence is indicated. Number of BlastX hits (E-value < 1e-10) against BmiGI and AvGI databases are shown.

Sequence ID	No. ESTs	Length (bp)	% GC	Longest ORF (aa)	Hits in BmiGI	Hits in AvGI
TC3	93	3236	41.04	115	2	1
TC1286	38	1912	40.48	90	2	0
TC4	28	2495	37.52	72	2	0
TC1357	20	1046	42.64	91	2	1
TC71	20	1529	41.86	112	1	0
TC91	16	1417	46.01	118	3	0
TC46	16	1049	48.4	113	2	1
TC96	14	1013	39.1	111	1	0
TC1313	14	995	49.05	115	2	1
TC109	13	2392	39.423	92	2	0

Table 4.2 Summary of properties of 10 out of 390 tentative consensus sequences (TC) containing no Open Reading Frames (ORFs) more than 100 amino acids (aa) in length. Location of the longest ORF within the TC sequence is indicated. Number of BlastX hits (E-value < 1e-10) against BmiGI and AvGI databases are shown.

Sequence ID	No. ESTs	Length (bp)	% GC	Longest ORF (aa)	Hits in BmiGI	Hits in AvGI
TC1286	38	1912	40.48	90	2	0
TC4	28	2495	37.52	72	2	0
TC1357	20	1046	42.64	91	2	1
TC109	13	2392	39.42	92	0	0
TC1441	10	1018	41.16	78	0	0
TC1288	9	1038	39.02	85	2	0
TC155	9	907	39	91	3	1
TC167	9	957	45.35	100	1	0
TC1467	8	976	44.98	92	1	0
TC1472	8	1308	43.88	87	5	0

© University of Pretoria

Table 4.3 Three tentative consensus sequences (TC) within RaGI containing no Open Reading Frames (ORFs) more than 50 amino acids (aa) in length. Number of BlastX hits (E-value < 1e-10) against BmiGI and AvGI databases is indicated.

Sequence ID	No. ESTs	Length (bp)	% GC	Longest ORF (aa)	Hits in BmiGI	Hits in AvGI
TC1864	3	441	34	40	1	1
TC794	2	222	43	46	0	0
TC2124	2	366	44	44	1	0

4.2.2 TC3-5 cluster

TC3, TC4 and TC5, comprising the cluster described in Chapter 3, have no significant database matches and lack long ORFs. Figure 4.1 shows a diagram of how the three sequences align with each other. TC3 and TC4 are 92.7% identical within the overlapping 2,631 bp region. TC3 and TC5 are 98% identical over 1,739 bp and TC4 and TC5 are 96.8% identical over 1,130 bp.

EST abundance and TC length is greatest in TC3 with 93 ESTs and 3,236 bp, followed by TC5 with 49 and 2,622 bp and TC4 constituting 28 ESTs and 2,495 bp long. The non-coding probability of TC3, TC4 and TC5 was predicted as 21%, 64% and 76%, respectively. The longest ORFs in TC3, TC4 and TC5 are 115, 72 and 148 aa, respectively. None of these ORFs have an in-frame ATG at the 5' end. The average G+C content of sequences in RaGI is 50%. In comparison, that of these three TCs ranges between 37% and 47%. This cluster is also present in the RA-NIH library, albeit not as abundantly transcribed as in RaGI. Sequences in this group have no significant hits to proteins in the GenBank NR database. ORFs in these sequences had no significant (E-value < 0.01) matches to conserved domains in the CDD database. In addition, homologues are absent in EST databases of all other tick. The overall properties of this complex transcript family suggest that these may well represent long non-coding rather than protein encoding RNAs and they are functionally important to *R. appendiculatus*.

© University of Pretoria

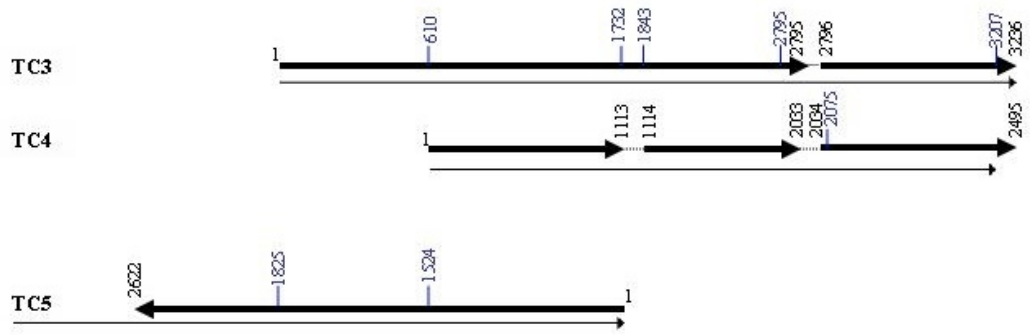


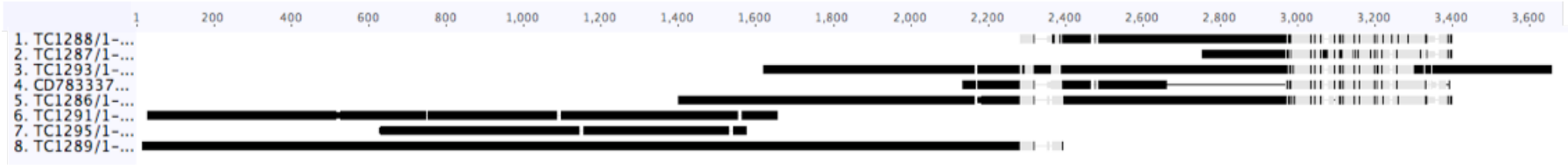
Figure 4.1 Diagrammatic representation of how sequences in cluster TC3-5 align relative to each other. Thin lines represent the entire sequence. Arrows represent aligned nucleotides with breaks indicating alignment gaps. Numbering above the arrows mark the nucleotide position of the sequence. Figure is to scale.

4.2.3 TC1286-93 cluster

The TC1286-93 cluster consists of seven TCs and one singleton, namely TC1286, TC1287, TC1288, TC1289, TC1291, TC1293, TC1295 and CD783337. The six TCs each have > 93% nucleotide identity when compared to one another. As also observed in the contigs comprising the TC3-5 cluster, none of the TCs in this cluster have a significant match within the GenBank NR database. TC1286 has the highest expression amongst the members of this group, containing 38 ESTs. All but TC1289 and TC1291 have an average G+C content of lower than the RaGI average of 50%. The mean length of sequences within the cluster is 1,254 bp ranging from 170 bp to 2,323 bp. The longest ORF of 285 aa occurs in TC1291, followed by TC1289, which has an ORF that is 216 aa in length. TC1286, TC1287, TC1288 and CD783337 have no ORFs longer than 100 aa. Nucleotide alignment of the members of this cluster is depicted diagrammatically in Figure 4.2.

© University of Pretoria

Figure 4.2 Diagrammatic representation of how sequences in cluster TC1286-93 align relative to each other. Solid lines represent aligned nucleotides, the breaks in the line indicating alignment gaps. Grey segments indicate weak similarity between sequences. Numbering above the arrows mark the nucleotide position of the sequence. Figure is to scale.



© University of Pretoria

A homologous cluster containing five sequences that are similar to TC1286-TC1293 in RaGI is present in BmiGI. It comprises one TC and four ESTs. All *R. appendiculatus* sequences in this group also have high sequence matches with unannotated mRNA sequences of the three-host tick, *D. variabilis* (nucleotide identity: 68% – 82%; E-value: 3e-05 - 2e-27). There are no homologues in the other tick gene indices. This might demonstrate conservation of ncRNA between different ixodid tick genera within the metastriate lineage.

4.2.4 Pseudogenes

Several sequences within RaGI were found to contain truncated copies of genes of known function in other organisms at either the 5' or 3' ends. In some of these transcripts the sequence adjacent to the protein-coding region contained no long ORFs. In certain cases, a full-length copy of the gene was also present in the EST database. Two such examples, TC1313 and TC1347, are discussed in Chapter 3, section 3.4.

Another example of a sequence that has sequence identity to part of a gene of known function is CD779866, which is 783 bp in length. Residues 1 to 62 of this singleton sequence have similarity ranging between 94-100% at nucleotide level and 84-100% at amino acid level, with the 5’ residues of the conserved protein ATP synthase c-subunit of tick species only, including the full-length copies of ATP synthase c-subunit in *R. appendiculatus*, *R. microplus* and *I. scapularis*. This N-terminal region is not conserved in other insects like mosquitoes, flies and lice. The functional domain of this protein lies between residues 75 and 132 of the ATP synthase c-subunit gene, which lies outside the region that aligns with CD779866. Nucleotides 188-783 of CD779866, which make up 76% of the sequence, have no significant similarities to any sequences in the GenBank database.

A further example of a partial match to a gene of known function is present within CD783970. It aligns at the 5’ 69 aa with the 3’ terminal of the *I. scapularis* phosphoserine phosphatase protein (GenBank accession: EEC06207.1) with 89% protein identity (in reverse orientation). The region of alignment to the *I. scapularis* protein contains a haloacid dehydrogenase (HAD)-like conserved domain (CDD:

© University of Pretoria

cd01427). TC1517 encodes a full-length copy of phosphoserine phosphatase (~230 aa), which aligns with 100% nucleotide identity (also in reverse orientation) to CD783970 over the 206 bp that code for the HAD-like domain. CD783970 is 821 bp in length and is sequenced from the 3' end of clone RAABA72. The EST sequenced from the 5' end of clone RAABA72, CD783971, has no sequence identity at nucleotide or protein level to known proteins in GenBank. If CD783970 represented a functional copy of the phosphoserine phosphatase protein we would expect its 5' counterpart, CD783971, to also contain sequence that was homologous to TC1517. The sequence downstream of the HAD-like domain in CD783970 contains a repetitive element found in several other RA sequences, including other singletons and sequences derived from the three randomly selected BAC clones that were sequenced (see chapter 6). One singleton containing this repeat element also contains a copy of Ruka, a SINE discovered as part of this study (detailed in Chapter 6), within the 5' flanking region. Given the evidence it seems likely that CD783970 represents a pseudogene of the phosphoserine phosphatase protein.

4.3 Conclusion

The characteristics of ncRNA in ticks have not been identified. The contents of the entire transcriptome of ticks is still unknown, therefore it is likely that some of the putative ncRNA transcripts identified above will have protein-coding functions. Having said that, the depth of coverage of transcripts in RaGI is not maximal (Chapter 3) so we are also likely to have missed identifying many more ncRNAs that may play important regulatory roles in tick biology. The limitations in *in silico* prediction of non-coding RNAs require that the putative ncRNAs be verified by laboratory experimental methods.

The validity of the above-mentioned sequences in which partial genes are associated with apparently non-coding RNA sequences should ideally be reconfirmed from ESTs derived from independent cDNA preparations to ensure that they do not represent artefacts. It is possible that the two 'hybrid' singleton sequences described above represent ancient pseudogenes, in which the 3' unannotated regions diverged very

© University of Pretoria

rapidly from the protein-encoding progenitor. It is also possible that they represent chimeras, however any chimeric sequences that were readily detectable were automatically discarded in the EST assembly pipeline (Quackenbush et al., 2000). An alternative explanation is that the non-coding region of the sequence homologous to its protein-coding transcript counterpart, such as that seen in the clusters TC1313-7 and TC1345-7 (Chapter 3), is a conserved regulatory region that acts on multiple genes, some of which are expressed in low copy numbers. As for CD783970, a very likely explanation could be that the singleton encodes a section of transposable element that has inserted itself within the protein-coding region thereby inactivating the gene (discussed further in Chapter 6).

© University of Pretoria

Chapter 5. Comparative analysis of gene indices generated from different ixodid tick species

5.1 Results

Transcriptomes of several tick species, both ixodid (hard) and argasid (soft) ticks, have been sequenced (Francischetti et al., 2009) in order to gain insights into their biology and identify predicted proteins that could be exploited for tick control. Here, I compare EST datasets from four ixodid tick species to identify transcripts that are conserved (1) between metastriate and prostriate lineages; (2) within the metastriate lineage; (3) within the genus *Rhipicephalus* and; (4) transcripts that are specific to *R. appendiculatus*. This information could be useful for understanding phenotypic differences between species, including feeding, life history strategies, host preference and vector-host interaction. This information will ultimately underpin the development of improved control strategies that are effective in controlling multiple species of ixodid ticks.

The EST libraries that were used to construct the gene indices for *R. appendiculatus* and *A. variegatum* (Nene et al., 2002; Nene et al., 2004) were derived from salivary gland tissues of fed female ticks, while those for *R. microplus* (Guerrero et al., 2005) were derived from multiple tick developmental stages and organs, that had been subjected to various treatments. The *I. scapularis* libraries were also generated from multiple tick developmental stages and tissues (Ribeiro et al., 2006; Valenzuela et al., 2002a). Size selection was performed prior to construction of all libraries from which the sequences on which the gene indices were based were generated. Only the BmiGI libraries were normalized.

© University of Pretoria

Following initial assembly approximately 26% of ESTs in RaGI were not assembled into TCs, compared to approximately 9% of *I. scapularis,* 11% of *R. microplus* and 42% of *A. variegatum* gene indices (Table 5.1). Tentative annotations have been assigned to 35% of sequences in RaGI, 34% in BmiGI, 39% in AvGI and 66% in IsGI.

Table 5.1 Summary of characteristics of gene indices for Ixodid ticks species

	RaGI	**BmiGI**	**AvGI**	**IsGI**
No. ESTs	18422	42651	3992	192746
% ESTs in TCs	74%	90%	59%	91%
% Annotated	35%	34%	39%	66%
Avg. TC length	858 bp	1008 bp	828 bp	1196 bp
Avg. % G+C	50%	47.6%	52%	51%

The high coverage of transcribed sequences determined from *I. scapularis* (192,746 sequences) and *R. microplus* (42,651 sequences), resulted in clusters incorporating a large proportion of the ESTs, and a relatively small percentage of singleton sequences. By comparison, the *R. appendiculatus* (18,422 sequences) and *A. variegatum* (3,992 sequences) EST datasets are small, although unlike *I. scapularis* and *R. microplus,* they are derived from only salivary gland transcripts and not a combination of expressed sequences from multiple tissues. The limited dataset explains the observation that a higher percentage of sequences within them did not assemble into clusters, despite the fact that these latter two libraries were not normalized prior to sequencing, while the IsGI and BmiGI libraries were normalized. Therefore, it was observed that the higher the number of transcribed ESTs sequenced, the higher the percentage of ESTs clustering into TCs. The *R. appendiculatus* and *A. variegatum* libraries were both derived from a single tissue - the salivary gland - at one time point, rather than from pooled tissues from different life cycle stages as is the case for *I. scapularis* and *R. microplus* libraries. Wang et *al*. (2007) hypothesize that there is a threshold beyond which additional transcripts sequenced do not bring about a significant increase in the number of TCs (Wang et al., 2007). Given that additional transcripts were identified through sequencing of a second library from the same RNA material (see later in this chapter for details), it is clear that the number of sequences in the *R. appendiculatus* salivary gland dataset is well below this threshold. However, it is difficult to speculate, regarding the

© University of Pretoria

number of ESTs that would be required to define the entire transcriptome, given that the size of the *R. appendiculatus* genome has yet to be determined.

Homologues of each *R. appendiculatus* TC sequence were sought in the *R. microplus* Gene Index (BmiGI) and *A. variegatum* Gene Index (AvGI) EST databases using the BlastN algorithm. BlastN was used so as to capture both protein-coding and non-protein coding transcripts that are conserved in the four ixodid tick species. The significance of non-coding RNA is discussed in Chapter 4. By using stringent cut-off values (BlastN E-value < 1e-25, combined with an alignment length > 300 bp) the probability of false positives arising from partial matches and repeat-rich regions was minimised.

R. appendiculatus is closely related to *R. microplus* (Barker and Murrell, 2004). *R. microplus*, previously *Boophilus microplus*, has recently been incorporated into the genus *Rhipicephalus*, based on comparative analysis of morphology and nucleotide sequences (Murrell and Barker, 2003). Homologues of *R. appendiculatus* gene sequences exist in all the three of the metastriate tick species (Av, Rm, Ra). Searching the entire RaGI dataset resulted in identification of 2,313 (32% of the total) RaGI sequences that exhibited similarity to 2,199 sequences in BmiGI (80% of the BmiGI matches were assembled into TCs). In addition, four hundred and sixty two (6%) of RaGI sequences exhibited similarity to 1,232 sequences within AvGI (16% TCs), and 667 (9%) RaGI sequences were similar to 1,880 sequences in IsGI (53% TCs) (Table 5.2). The large differences observed in the number of sequences that were similar between RaGI and AvGI (462 vs 1,232) and RaGI and IsGI (667 vs 1,880) was a result of one of the following: 1) a higher number of transcripts encoding paralogous gene families where one RaGI sequence exhibited similarity to more than one distinct transcript in the other GI with comparable Blast scores; 2) incomplete sequencing of full length mRNA demonstrated by matches to two or more distinct sequences over consecutive regions within a single RaGI sequence; 3) insufficient numbers of transcripts sequenced, resulting in one RaGI sequence exhibiting identity to several unassembled singletons in the GI that was being compared over different regions. Poor sequence quality or incorrect assembly could also contribute to low quality alignment,

© University of Pretoria

although the extent of this is difficult to assess. The above scenarios give a glimpse into the composition of an assembled EST library.

Table 5.2 Number of sequence matches between RaGI and the gene indices of R. microplus (BmiGI), A. variegatum (AvGI) and I. scapularis (IsGI).

Gene Index	No. RaGI sequences similar in the gene index	No. sequences in the gene index similar in RaGI
BmiGI	2312	2199
AvGI	462	1232
IsGI	667	1880

Interestingly, a larger proportion of unannotated compared to annotated RaGI sequences (TCs and singletons) have matches in BmiGI and AvGI - 35% and 65% in BmiGI, 40% and 60% in AvGI, when comparing annotated and unannotated, respectively. In contrast, more annotated (70%) than unannotated RaGI sequences have matches in IsGI. The likely explanation is that, as mentioned above, *R. appendiculatus* is evolutionarily considerably more distant to *I. scapularis* than it is to *R. microplus* or *A. variegatum*, and this is reflected in the conservation of genomic sequences between the species as sequences conserved between distant species are more likely to be annotated. The relatively high percentage of conserved unannotated sequences present in the two different genera *Rhipicephalus* and *Amblyomma* is noteworthy and suggests that the unannotated sequences may play a functional role, despite their lack of database matches. However, this hypothesised functional conservation of unannotated sequences is not apparent when the more distantly related metastriate and prostriate lineages are compared. The stringent search criteria used may have resulted in the reduced number of matches observed between the more divergent metastriates and prostriates. The trans-lineage conserved sequence matches tend to have functional annotation, presumably due to the constraints imposed by encoding proteins. The distribution of annotated and unannotated RA-NIH sequences matching BmiGI and AvGI is opposite to that observed for RaGI, i.e. a higher proportion of annotated compared to unannotated RA-NIH sequences had similarity with sequences in the gene indices - 57%, 80% and 86% of annotated RA-NIH sequences having matches in BmiGI, AvGI and IsGI, respectively. This is surprising, as the ratio of annotated to unannotated sequences in both RaGI and RA-NIH libraries is roughly equal, the proportion of

© University of Pretoria

annotated/unannotated hits to the gene indices would also be expected to be roughly equal. A high count of GI matches with the 124 new annotated NIH transcripts (see Chapter 3) could result in a comparatively higher proportion of annotated sequence hits in the NIH library however, this was not found to be the case as 24%, 3% and 19% of the annotated matches in BmiGI, AvGI and IsGI, respectively, were new annotated NIH transcripts, proportions that are not large enough to account for the difference in values observed between the two *R. appendiculatus* libraries.

The proportion of matches in the gene indices that are annotated is close to 50% in both BmiGI and IsGI indicating that many sequences that lack tentative annotations (TA) in RaGI match sequences that have been tentatively annotated in BmiGI and IsGI. In most cases, this arises when the unannotated RaGI sequence has similarity to a non-protein coding region (putatively an untranslated region) adjacent to the protein-coding ORF that determines the TA for that sequence. Sequencing of additional clones from the *R. appendiculatus* EST library is likely to alter these percentages by extending the contig assembly into the coding regions of the *R. appendiculatus* transcripts. Nevertheless, the vast majority of the sequences (92%) that are unannotated and have matches in BmiGI match unannotated sequences in the latter database. This is probably most frequently a result of conservation of non-coding RNA sequences with regulatory functions, but in some cases may be derived from proteins that are unique to the metastriate lineage of ticks.

The average G+C composition of RaGI is about 50%. That of the unannotated and annotated TCs is 47.1% and 53.5% respectively. This is consistent with the expected greater level of constraint on the G+C content of annotated TCs, the majority of which are likely to be protein-coding, as a result of codon usage bias.

5.2 Sequences conserved in ticks

Sequences in the four gene indices (RaGI, BmiGI, AvGI and IsGI) were compared to identify those transcribed in all the four ixodid tick species. The nucleotide-based

© University of Pretoria

search, BlastN (E-value < 1e-25, alignment > 300 bp) was used in order to speed up the search and to capture conserved non-protein coding transcripts. Although transcription data for other tick species is available in GenBank, I limited the comparative dataset to the species in the gene indices in order to standardise on one transcriptome assembly algorithm.

Only 147 sequences (TCs and singletons) were found to be present in all the four tick gene indices. Of these, 135 (92%) were TCs and 12 were singletons in RaGI. One hundred and forty four of the 147 sequences had assigned tentative Gene Ontology (GO) annotations based on similarity hits of E-value < 1e-25 to the GO database. Based on the tentative annotations (TA) assigned in the gene index and BlastX matches to known proteins in the non-redundant database where the TA was absent, all but four of the conserved RaGI TCs are putative housekeeping genes. One of these four non-housekeeping genes has a serine protease inhibitor domain (discussed in detail in chapter 3), while a second one (TC54) shows ~40% amino acid sequence similarity to proteins in arthropods that are members of the immunoglobulin superfamily (IgSF), including Basigin (BSG) precursor, Hemicentin-like and Neuroplastin-like proteins, some of which are involved in cell recognition, binding or adhesion. Of particular interest is BSG-precursor, which has been identified as a receptor that is located on erythrocytes and is essential for *Plasmodium falciparum* cell invasion (Crosnier et al., 2011), inhibition of BSG having been demonstrated to block cell invasion. TC54 contains 24 ESTs indicating that the protein is expressed at a fairly high level. Protein fold prediction of the 269 aa ORF encoded by transcripts within TC54 using GenThreader exhibited high scoring (score >65) structural similarity to certain T-cell receptors (PDB ID: 1HNF, 1SY6) suggesting a possible role in immune evasion. It contains a predicted signal peptide with a cleavage site between residues 20 and 21, indicated it is likely to be secreted. Presence of BSG precursor in several tick as well as arthropod species indicates the transcripts are not contaminants. Furthermore, transcripts constituting TC54 were derived from both the infected and uninfected libraries. The results obtained using BSG precursor, in which inhibition blocked host erythrocyte invasion by *P. falciparum,* suggest that the TC54 molecule is worthy of further investigation in the

© University of Pretoria

context of application for disease control. However, it's function in arthropods remains unclear.

Two TCS (TC1112 and TC169) have similarity over more than 175 aa with sequences of unknown function (hypothetical proteins) present in many arthropods, including other ticks, lice, ants, flies, butterflies, fleas and also in mammals including humans, mice and monkeys. TC1112 has an ORF of 200 residues that contains a match to the conserved domain NOA36 (pfam06524) (E-value: 9e-99). The function of NOA36 is unknown, however it contains 29 Cysteine residues that are conserved at specific positions. Cysteines are commonly found in enzyme active sites. The side chains of cysteine residues covalently link to form a disulphide bridge. All the proteins that match TC1112 in the GenBank NR database contain the NOA36 domain. Some of these proteins carry the annotation 'zinc finger protein 330-like protein' (Uniprot: Q9Y3S2), which, based on its nucleolar localization, is thought to be involved in regulation of the cell cycle and transcription (Bolívar et al., 1999). TC169 has an ORF of 175 residues, which contains the DUF1077 conserved domain (CDD:191513), which is found in proteins of unknown function. This domain is present in all proteins in GenBank NR that match to TC169. Some of these matches are annotated as 'transmembrane protein 85', which is thought to mediate anti-apoptotic activity (Uniprot: Q5J8M3).

Within the RA-NIH dataset 28 sequences had high confidence matches (E-value < 1e-25 and alignment length > 300 bp) to the GIs derived from all three of the other tick species. All but one of the 28 is present in RaGI.

Given the evolutionary distance between *R. appendiculatus* and *I. scapularis* it can be assumed that the conserved sequences identified above are conserved in most, or all, ixodid tick species.

© University of Pretoria

5.3 *R. appendiculatus*-specific transcripts

Sequences present in RaGI that do not have similarity to sequences in any database, at either nucleotide or protein levels, total 250 TCs. These sequences currently appear to be unique to *R. appendiculatus*. However this is limited by the level of coverage of tick and other arthropod databases and particularly the tissue pooling strategy utilised for the construction of BmiGI, which limits the scope for direct comparison with RaGI.

The protein-coding probability of each of the 250 'RaGI-unique' sequences was predicted using PORTRAIT. More than half the sequences are predicted to be protein coding of which 66 have a predicted coding probability of > 90%, 41 sequences have a coding probability from 90 –70%, and the remainder have a coding probability of between 70 and 5%. Of the 66 sequences having coding probability over 90% ORFs of 14 had significant matches to domains in CDD (E-value < 0.01). The results for the CDD search are summarized in Appendix Table A1.4.

Of the 250 sequences, currently identified only in RaGI, TC5 (see Chapter 4 for details) contains the highest number of ESTs (49 ESTs) followed by TC1354, which is assembled from 23 ESTs. TC1354 is 1,685 bp in length and has a coding probability of 99.53%. Four ORFs longer than 100 aa are present in the TC, the longest of them being 166 aa, beginning with an in-frame Met residue and closest to the 5' end of the sequence. The ORF contains an 80 aa GIY-YIG endonuclease domain (CDD search: E-value = 6e-07; % identity = 33%; superfamily Accession: cl15257) which is present in many proteins involved in cellular processes, as well as in eukaryotic transposable elements. The domain consists of two motifs, GIY-YIG and CCHH motifs (Figure 5.1). The latter motif is a distinct characteristic of the Penelope-like transposable element (CDD Accession: cd10442) that is widely found in eukaryotes. When searched using GenThreader and HHPred methods none of the ORFs in TC1354 matched the structural profiles of known proteins with significant values. The presence of an endonuclease motif together with the lack of sequence similarity with known proteins suggests TC1354 has a high

© University of Pretoria

probability of coding for a retrotransposon. Transposable elements are discussed in further in Chapter 6.

```
                  #          #                                  #      #
TC1354   44 VLYEVPFECGKRYIGQTGRCVNDRLREHRYNVGRAQRDPSGSYGTLASHSVSCC-EPDFN 102
TC1354 103 RTRILARGVHDTQYRRWVE 121
```

Figure 5.1 The GIY-YIG endonuclease domain (cd10442) in TC1354 identified using protein sequence search against the conserved domain database (CDD). GIY-YIG motif is in bold. CCHH conserved motif found in Penelope-like elements was also identified (residues underlined and indicated by a hash). As the role of CCHH motif is unknown the significance of the absence of second Histidine in the motif is yet to be determined.

Both TC5 and TC1354 are examples of transcribed sequences, which are abundantly expressed, yet cannot be assigned a function using current similarity and prediction-based algorithms. This is consistent with the hypothesis that they encode *R. appendiculatus*-specific proteins. Although as additional data accumulates, particularly from more closely related tick species within the genus *Rhipicephalus*, these TCs may ultimately prove not to be *R. appendiculatus*-specific. However, the existence of such highly transcribed sequences, with predicted coding potential highlights the limited state of current knowledge of arthropod, and particularly ixodid tick, genomics.

TC889, which is 1,038 bp in length and composed of two ESTs, has the highest predicted coding probability of all the '*R. appendiculatus*-specific' transcripts (99.96%), yet none of its three ORFs longer than 100 aa have significant matches using structure profile searches. The sequence also has three ORFs between 50 and 100 aa long. No significant primary sequence similarities were found for these short ORFs in the GenBank NR database using the BLAST suite of programs. Searches for distantly related sequences followed by application of secondary structure similarity prediction software using HHPred also did not result in matches with significant E-value scores. However, it is interesting to note that for all the three short ORFs, matches with high E-values (E-value from 2 to 12), which would typically not be considered significant, were to immunity-related proteins such as defensin, agglutinin and interleukin-8. Having observed a finite, but low PORTRAIT prediction error rate of < 5% (Chapter 2) it is possible (although not

© University of Pretoria

likely) that the prediction that this transcript is protein encoding is incorrect and that it is actually non-coding RNA.

The longest ORF (potentially encoding 441 aa) amongst the unique *R. appendiculatus* transcripts is located in TC796. This TC is composed of a pair of ESTs derived from a single clone from the infected salivary gland EST library. The G+C content of the ORF is 53% and it contains a predicted signal peptide. According to localization prediction using PSORTII (Horton and Nakai, 1997) the ORF is likely to be extracellular. An ORF this large is likely to contain multiple domains. Functional domain prediction using pDomTHREADER resulted in a high confidence match (score: 5.895; P-value: 1e-05) with the structural domains of B19 parvovirus capsid protein VP2 (PDB: 1S58A00).

Searches for conserved structural folds using the pGenThreader algorithm (Lobley et al., 2009) located within a 245 aa ORF encoded by TC1976 predicted a fold resembling bacterial soluble lytic transglycosylase (SLT) (PDB ID: 1QSA A0) with high confidence values (P-value 2e-04; Score 54.046). SLT is a lysozyme that hydrolyses bacterial cell wall. In ticks it might have antimicrobial properties.

Of the sequences obtained from the NIH library (whose construction and analysis is described in Chapter 3) that are not present in the original RaGI database, 92 have no sequence matches to known proteins in the GenBank NR (E-value < 1e-2). All 92 sequences are singletons of median length 337 bp, ranging between 59 and 1103 bp. Seven of these sequences have an ORF longer than 100 aa while 40 sequences have no ORFs longer than 50 aa. Searches for conserved structural domains of all ORFs longer than 100 aa using three different algorithms (GenTHREADER, pGenThreader and HHPred) resulted in no significant similarities to known protein folds. All but two of these ORFs have a high non-coding probability (> 70%).

© University of Pretoria

5.4 Conclusions

A higher proportion of unannotated sequences were similar between *R. appendiculatus* and *R. microplus* (65%)/*A. variegatum* (60%) than between *R. appendiculatus* and *I. scapularis* (30%). By contrast a large proportion of annotated sequences were similar between *R. appendiculatus* and *I. scapularis*. This is consistent with the reported evolutionary distances between the metastriate and prostriate lineages if it is assumed that non-coding sequences diverge more rapidly over long periods of evolutionary time.

I found 127 sequences to be conserved in all four tick species. A majority of them encoded putative housekeeping genes. More interesting were the small number of universally conserved sequences without a currently known function. One of these is likely to be involved in immune evasion. The number of sequences found to be conserved in all four ticks is greatly under estimated by the limitation in EST dataset size. Given that 667 RaGI sequences were conserved with *I. scapularis*, which is evolutionarily the most distant to *R. appendiculatus* of the four tick species studied here, it can be expected that the as yet unidentified homologues of these 667 sequences in BmiGI and AvGI are also conserved. 30% of these RaGI-IsGI conserved sequences have no assigned annotations highlighting an opportunity for discovery of novel tick-specific proteins.

I observed that 250 TCs are specific to *R. appendiculatus.* These sequences did not include singletons, which account for 24% of the total ESTs, however, a detailed analysis of these TCs using coding region prediction and structural algorithms provides convincing evidence that among highly expressed *R. appendiculatus*-specific sequences there are a number of novel protein coding sequences that may be confined to *R. appendiculatus* and its close relatives.

© University of Pretoria

Chapter 6. Analysis of the nuclear genome of *R. appendiculatus*

Tick genomes are more similar in size and organization to those of vertebrates than the streamlined genome of the model arthropod *Drosophila melanogaster.* Genome size estimates for ticks, based on re-association kinetics, range from 2 x 10^9- 7 x 10^9 (Guerrero et al., 2010; Ullmann et al., 2005). Tick genomes contain on average ~25% highly repetitive (HR) DNA, ~40% moderately repetitive (MR) DNA and ~34% unique DNA (Ullmann et al., 2005). HR DNA includes low complexity and satellite DNA while MR consists of transposable elements and multigene families. A high proportion of the repetitive component of the genomes of *I. scapularis* and *R. microplus* is arranged in the short period interspersion pattern of organization, which is characteristic of the majority of animal species. In contrast, despite its large genome (1.04 x 10^9 bp), *A. americanum* repetitive DNA exhibits long period interspersion, typically observed in organisms with small genomes (Ullmann et al., 2005). As described in earlier chapters some of these genomes contain non-coding sequences. In this chapter, I analyse a subset of the *R. appendiculatus* genome to investigate the proportion of single copy and repetitive DNA components present, in particular the classes of transposable elements that comprise the MR component of the genome. Through mapping of ESTs onto the corresponding genomic loci I also attempt to understand the structure and organization of genes in this species.

© University of Pretoria

6.1 Insights into the organization of the *R. appendiculatus* genome through analysis of sample sequences

Three randomly selected clones (RAHD, RAHE, RAHF) isolated from a bacterial artificial chromosome (BAC) library of *R. appendiculatus* generated from high molecular weight DNA by the commercial company Amplicon Express (WA USA) were sequenced at the institute for genomic research (TIGR) using a shotgun strategy based on sub-cloning into plasmids and conventional Sanger di-deoxy chain termination chemistry. The assembly was performed as described in Desjardin et al., 2007. None of the three BAC clones were assembled over the entire cloned sequence contained within it, most likely due to the presence of repetitive sequences, however, insufficient coverage of shotgun sequencing could also have been a contributing factor. Clone RAHD was assembled into three contigs (arbitrarily numbered 48, 50, 87), clone RAHE into two contigs (5, 71) and clone RAHF into six contigs (73, 74, 77, 78, 80, 81). The sizes of the contigs ranged from 2-90 kb and are listed in Table 6.1.

Additional sample genomic sequences from the *R. appendiculatus* genome were obtained from a separate study. This involved sequencing of total genomic DNA prepared from a Muguga cocktail *Theileria parva* sporozoite infection and treatment (ITM) vaccine stabilate using the next generation Roche 454 pyro-sequencing technique. The sequencing run generated approximately 15 Mb of raw sequence data from 116,937 reads of which 102,446 reads were singletons, 2,743 were repeats and 3,059 were < 50 bp in length. 9,443 reads assembled into 674 contigs made up of 174 Kb, representing a small percentage (1%) of the total data generated. The results of the assembly are summarized in Table 6.2.

© University of Pretoria

Table 6.1 Size of contigs obtained for three R. appendiculatus BAC clones (RAHD, RAHE, RAHF) sequenced using the Sanger method. The accession number for each contig submitted to GenBank is listed in column 3.

Clone	Contig	GenBank Accession	Size
RAHD	48	EU018129.1	65 kb
	50	EU018130.1	2 kb
	87	EU018131.1	85 kb
RAHE	5	EU018132.1	6 kb
	71[a]	EU018133.1 EU018134.1	90 kb
RAHF	77	EU018137.1	2 kb
	78	EU018138.1	9 kb
	80	EU018139.1	57 kb
	81	EU018140.1	6 kb
	74	EU018136.1	12 kb
	73	EU018135.1	21 kb

[a] Contig RAHE-71 contained an *E. coli* insert of 786 bp. Removal of the artefact from the contig resulted in two subcontigs.

Table 6.2 De novo assembly using gsAssembler, of ITM stabilate sequenced by Roche 454 pyrosequencing technique.

	Number of reads	Number of bases
Total	116937	15457231
Assembled	9443	174000
Singletons	102446	10535270
Repeats	2743[1]	400041
Length < 50	3059[2]	139524
Outliers	338[2]	63606

[1] Some reads were included in contig assembly
[2] Reads were not included in contig assembly

Within the assembled component, 171,616 bp (assembled into 659 contigs) were of tick origin and the rest originated from *T. parva*, or the mammalian host. ITM vaccine stabilate is generated from a supernatant derived from whole ground *T. parva* infected ticks. It is therefore not surprising that the majority of contigs in the stabilate were derived from the tick genome, particularly given the approximately hundred fold difference in genome size between the *T. parva* parasite and *R. appendiculatus* vector genomes.

© University of Pretoria

A total of 526.6 Kb of *R. appendiculatus* genomic sequence data comprising 356 Kb from the BAC clones and 171.6 Kb from the ITM sample was therefore available for analysis. In the case of the ITM stabilate genomic DNA assembly the singletons were excluded from analysis because the average read length for the singletons was too short (132 bp, range: 50–529 bp) to yield meaningful results. Therefore, only the genomic sequences that had been assembled into contigs were used. Although representing less of the overall sequence dataset, because most of the contigs were short (Average length = 258 bp, range 99-1194 bp) we assume that the ITM sequences provide a more representative indication of the *R. appendiculatus* genome overall, since according to the methodology used to create the library, they should have been derived from randomly distributed locations.

I initially attempted to ascertain if any of the expressed sequences within RaGI could be mapped back onto the genomic sequences. This would not only help to annotate the genomic sequence, but would also, for the first time provide insight into the structure and organisation of specific genes in *R. appendiculatus*. For this exercise a combination of BlastN and the gmap algorithm (Wu and Watanabe, 2005) were employed using default settings. Although ~31 Kb of sequences derived from 128 RaGI sequences had significant matches (>70% identity over >200 bp) to genomic regions (71 to sequences within BAC clones and 57 to the sample ITM data) no RaGI sequence could be mapped onto the genomic sequences over their entire length. If, by analogy with other ixodid tick species, the genome size of *R. appendiculatus* is assumed to be 1 x 10^9bp (which is probably a conservative estimate) the BAC sequences comprise only 0.000355% of the genome. It is therefore, not surprising that none of the EST data matched 100% over the entire length of the transcripts with the sequences cloned within the BACs. However, the partial EST matches suggest considerable redundancy within the *R. appendiculatus* genome with many closely related loci presumably present.

Among the RaGI sequences that matched the genomic sequences, 41 contained repeat elements and mapped to multiple genomic loci. They additionally aligned with several other RaGI transcripts with 60-70% identity, in regions corresponding to those that

© University of Pretoria

mapped to the genomic sequences. Of the remaining 87 non-repeat sequences that matched the genomic sequences 29 were TCs and 58 (67%) were singletons. Sequence searches of the 29 TCs against GenBank nucleotide and protein databases resulted in similarities to sequences annotated as mitochondrial gene Cytochrome c Oxidase I, exonuclease proteins, the integrase core domain commonly found in retrotransposons, vertebrate tubulin-specific chaperone C-like proteins and ribosomal RNA genes. Nuclear-encoded rRNAs are known to lack mature 3' end polyA tails since they are transcribed by RNA polymerase I, therefore their presence in the EST database was intriguing. rRNA transcripts having poly adenylated tails at various positions along their length have been identified in eukaryotic organisms as diverse as the unicellular baker's yeast *S. cerevisiae* to the *Homo sapiens*. These transcripts are believed to be degradation intermediates (Slomovic et al., 2006). A total of 48 ESTs in RaGI encoded rRNA. Some transcripts however have a high probability of being rRNA pseudogenes functionally inactivated by the insertion of a retrotransposon. These transcripts are described in detail in section 6.2.2 below.

6.2 Transposable element-like sequences in *R. appendiculatus*

Transposable elements are segments of DNA with the capacity to transpose (i.e. move) between non-homologous sites in the genome. Classified on the basis of their mechanism of transposition, transposable elements fall into two classes. Class I transposable elements rely on retrotransposition, which uses reverse transcriptase for genomic integration, and are often known as retrotransposons. Class II elements, also known as DNA transposons, transpose using a cut-and-paste method or a rolling-circle mechanism (Kidwell, 2002). Class I transposable elements include long terminal repeat (LTR) retrotransposons, which include the Ty1 and Ty3 identified in *S. cerevisiae* and the Copia and Gypsy-like elements of *Drosophila*. Non-LTR retrotransposons include long interspersed nuclear elements (LINEs), for example L1 in humans and R1/R2 in *Bombyx mori*, and short interspersed nuclear elements (SINEs), for example the Alu

© University of Pretoria

family in humans. Examples of Class II DNA transposons include hAT in mouse, Hobo and P in *Drosophila*, Tc1 in *Caenorhabditis elegans* and Helitrons in *Arabidopsis thaliana*.

6.2.1 The Ruka SINE element

BlastN searches of the BAC contigs against each other revealed the presence of 28 copies of a repetitive region ~240 bp in length, randomly distributed within the BAC contigs from clones RAHD and RAHF. The repeat sequences comprised approximately 2% of the 242 Kb of genomic sequence represented in the contigs of the two clones. When searched against the GenBank database the repeat sequences were found to be most similar (E-value ≤ 1e−35; Identity ≥ 85%) to the genomic sequence within intron 7 of the glucose-6-phosphate dehydrogenase (G6PDH) gene from *R. microplus* (GB Accession: DQ118973). As this gene has been well characterized its version of the repeat sequence, which is more likely to be full-length, was used as a reference in subsequent analyses. The repeat sequence was designated 'Ruka', from the Kiswahili word meaning 'to jump'.

The Ruka sequences presented characteristics of a tRNA-derived SINE (Figure 6.1). Direct repeats, 10–14 bp in length, flanked the sequence. The A and B boxes of the polymerase III promoter (in boxes in Figure 6.1) and a tRNA-like region (underlined in Figure 6.1) were present on the 5' end. The 3' end contained poly-pyrimidine tracts (in green in Figure 6.1) and was relatively A-rich compared to the rest of the Ruka sequence. A Polymerase III terminator, TTTT, at the 3' end was found in most, but not all of the, Ruka sequences.

© University of Pretoria

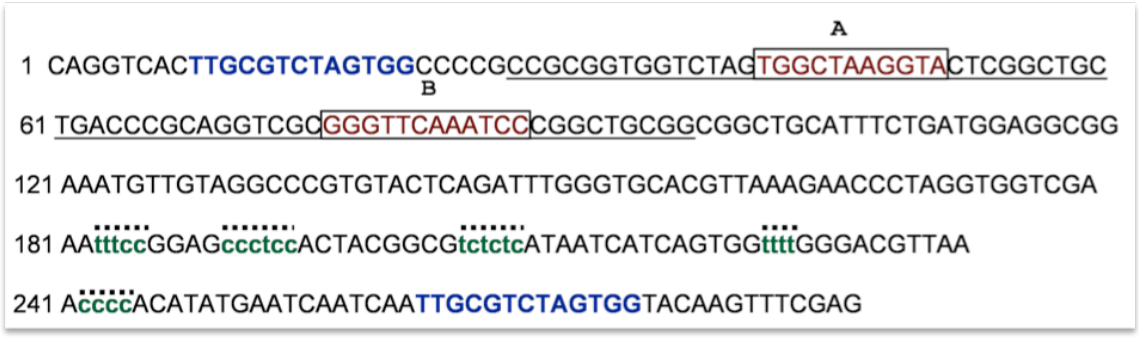


Figure 6.1 Characteristics of Short Interspersed Nuclear Elements (SINE) identified in Ruka-SINE sequence from R. appendiculatus. The Polymerase III promoter boxes A and B are enclosed in boxes; a tRNA-related sequence is underlined; poly-pyrimidine tracts are in green bold lower case text and under dashed lines; short direct repeats are indicated by blue bold capitalised text.

tRNA secondary structure prediction of the tRNA-like sequence at the 5′ end of Ruka using tRNA Scan-SE showed the sequence folded into a pseudo-tRNA structure similar to that of a serine tRNA, as determined by the GCU anticodon presented in the fold, which recognises and incorporates the amino acid serine. This is in contrast to a lysine tRNA more typically found in SINEs (Shedlock and Okada, 2000), which contain a UU[U/C] anticodon. Although there is no confirmed explanation for the higher prevalence of lysine tRNA-derived SINEs observed in most organisms, it is thought that this structure binds more effectively to reverse transcriptase and may therefore multiply more efficiently (Okada, 1991).

BlastN search of the *R. microplus* Ruka sequence from the G6PDH intron against RaGI (with the low complexity filter disabled) gave matches to 49 sequence (E-value < 1e−10). Thirty-four of these were 180 bp or longer. The characteristics of a serine tRNA-derived SINE were present in all RaGI matches.

The Ruka transcripts that I identified were probably generated by RNA Pol II, since the libraries from which the EST gene indices were constructed were derived from polyadenylated transcripts and Pol III transcripts are rarely polyadenylated (Düvel et

© University of Pretoria

al., 2003). The method of priming for EST gene index library construction depended on the template sequences being polyadenylated. Eleven random Ruka-containing clones from the RaGI cDNA library were sequenced, using vector primers and Ruka-specific internal primers. In ten of the 11 clones it was confirmed that the sequences were polyadenylated demonstrating that these were genuine transcripts and not the result of genomic DNA contamination of the library. The Ruka sequences found in RaGI and BmiGI are therefore very unlikely to be intermediates in the transposition process but are more likely indicative of promiscuous Pol II transcription.

6.2.1.2 Presence of Ruka-like elements in tick gene indices

BmiGI, AvGI and IsGI were searched (BlastN) using the reference G6PDH intron 7 Ruka sequence of *R. microplus* to identify sequences related to this element. Similar sequences with an E-value ≤ 1e−10 were found in all the databases (Table 6.3).

Table 6.3 Frequency of occurrence of Ruka-like sequences within gene indices of four ixodid tick species (RaGI, BmiGI, AvGI, IsGI), and BAC sequences of R. appendiculatus. Ruka sequence from the R. microplus glucose 6-phosphate dehydrogenase intron was queried against the databases. Sequence matches with E-value ≤ 1e−10 and nucleotide identity ≥ 60% were selected.

Species	No. homologs	[1]Ali.Len. range	% identity range
BmiGI	37	181–244	66–93
RaGI	34	180–242	64–86
R. appendiculatus BAC	8	189–231	84–87
AvGI	7	199–251	64–78
IsGI	3	208–242	68–70
BmiGI	23	80-168	66-95
RaGI	16	99-179	70-91
R. appendiculatus BAC	20	53-172	83-96
AvGI	3	106-171	71-84
IsGI	0	n/a	n/a

[1] ali. len = alignment length;

More than 30 Ruka-like sequences of length 180 bp or more and with > 80% nucleotide identity were found in both the BmiGI and RaGI databases. The AvGI database contained seven sequences that had sequence identity to Ruka over > 180 bp, while three IsGI sequences exhibited approximately 70% identity over > 180 bp. The Ruka sequences,

© University of Pretoria

longer than 180 bp, identified in the RA BAC assemblies, RaGI and BmiGI, were aligned and an unrooted cladogram, shown in Appendix Figure A2.1, was generated using a tree searching maximum likelihood algorithm. The sequences clustered into two major groups. One cluster predominantly comprised sequences from *R. microplus*, whereas the second was a mixture of sequences from both *R. microplus* and *R. appendiculatus*. When the analysis was repeated using *I. scapularis* as an outgroup the results remained similar (Appendix Figure A2.2). I presume that given the very close phylogenetic relationship of *R. appendiculatus* and *R. microplus*, which have recently been re-classified in the same genus (Barker and Murrell, 2004) that the Ruka sequences in cluster two may have been present in the common ancestor of these two tick species.

Due to the irreversible, random nature of their insertion, SINEs have been used as a molecular tool for examining phylogenetic relationships (Shedlock and Okada, 2000). Twenty conserved sequences containing Ruka from each of the four tick GI – RaGI, BmiGI, AvGI and IsGI – were queried against each other (E-value ≤ 1e-10) to obtain the versions of Ruka whose flanking regions are conserved in all four ticks. Conserved flanking regions would indicate that the Ruka copy had inserted at that locus before the two species diverged. Four loci with sequence similarity in the regions flanking the Ruka insertions were identified (Table 6.4). Two of these were common to *R. appendiculatus*, *R. microplus* and *A. variegatum,* one was conserved only in *R. appendiculatus* and *R. microplus* while one was common to *R. appendiculatus*, *R. microplus* and *I. scapularis*. The latter Ruka insertion loci had the same tentative annotation (nucleoside di-phosphate kinase) in all three species.

Table 6.4 Conserved Ruka insertions at common loci shared between two or more tick species

Locus	***R. appendiculatus***	***R. microplus***	***A. variegatum***	***I. scapularis***
1	CD787372	TC1925	BM292626	No match
2	TC1188	TC57	No match	ISCW023074-PA
3	CD789280	TC3432	BM290387	No match
4	TC1552	CK191250	No match	No match

The identifier for the tentative consensus sequence (TC) and/or singleton (CD/ CK/BM) is provided for each species.

© University of Pretoria

These loci are likely to be the sites of ancestral insertions that occurred prior to the evolution of different ixodid genera and speciation within the genus *Rhipicephalus*. It is not surprising to find that three loci were not conserved in *I. scapularis* since the prostriate lineage, which contains *I. scapularis* and other species within the genus *Ixodes*, is believed to be evolutionarily highly divergent from the other three tick species that are classified within the metastriate lineage, according to molecular phylogenetic markers, although there is no well defined evolutionary timescale for this split (Barker and Murrell, 2004, Fig. 1). The lack of conservation of sequences flanking the Ruka insertion sites provides additional evidence supporting an ancient split. Given the molecular divergence of the *Ixodes* prostriate lineage (Barker and Murrell, 2004), from the metastriate some elements within the ancestral Ruka transposable element family are likely to have undergone significant mutational changes in *I. scapularis* resulting in the creation of novel SINEs that cannot be detected using sequence similarity-based comparisons, although future secondary structure analyses may reveal additional structurally conserved copies. The Ruka elements observed in both *I. scapularis* (three copies) and *A. variegatum* (10 copies) may represent 'master copies' that are more highly represented due to their frequent transposition (Kidwell, 2002).

6.2.1.3 Quantification of Ruka copy number in the *R. appendiculatus* genome using real time PCR

Six primer pairs (described in Section 2.6.3 in Methods) derived from the eight most conserved Ruka copies in the *R. appendiculatus* BAC sequences were used to amplify Ruka sequences from the genomic DNA of *R. appendiculatus* using real time PCR (RT-PCR). The results are summarised in Table 6.5. An average of between 5,100 and 28,800 genomic copies of Ruka were computed by the RT-PCR software following amplification using the six Ruka primer pairs (Table 6.5). In order to provide an indication of the probability of a Ruka primer pair amplifying a specific copy of Ruka, the percentage similarity of the Ruka primer sequences to all the copies of Ruka in the RA BAC sequences sequenced was calculated. A good match was defined as 100% identity in two or more bases at the 3' end, combined with the entire primer sequence exhibiting

© University of Pretoria

>80% similarity with the relevant section of the cloned genomic copy (Table 6.5, column 5). This revealed that between 75% and 88% of the Ruka copies in these clones would theoretically be amplified by individual Ruka primer pairs 1–6.

Table 6.5 RT-PCR using R. appendiculatus genomic DNA as a template with RUKA primer pairs 1–6.

Primer pair	R^2	PCR efficiency	[1]Average copy no.	[2]% Ruka amplifiable	[3]Predicted copy no.
Ruka 1	0.997	0.67	1.05E+04	75	1.40E+04
Ruka 2	0.997	0.78	5.11E+03	88	5.81E+03
Ruka 3	0.993	0.92	2.02E+04	88	2.30E+04
Ruka 4	0.992	0.74	2.88E+04	75	3.84E+04
Ruka 5	0.991	0.84	1.45E+04	88	1.65E+04
Ruka 6	0.995	0.88	1.14E+04	88	1.30E+04

[1] per haploid genome, [2] in RA BAC, [3]Predicted copy no. in haploid genome

Assuming that (1) *R. appendiculatus* has a genome size of 1×10^9 and (2) these frequencies at which Ruka occurs in the RA BAC sequences are representative, then a total of 65,000 copies of this repeat would be predicted as being present in the *R. appendiculatus* genome. The copy numbers indicated by the quantitative-RT-PCR (qRT-PCR) results were almost certainly minimal estimates as 1) there was lack of sequence conservation in some Ruka copies and 2) it is possible that there was primer competition in the genomic DNA. Due to uncertainty of the efficiency with which a specific primer pair would amplify multiple variant copies of Ruka in the *R. appendiculatus* genome it was difficult to provide a more precise estimate using this methodology. Given these factors, the agreement of the qRT-PCR data with the copy number extrapolation based on limited genomic sample sequencing is surprisingly good.

6.2.2 The presence of other Class I transposable elements in *R. appendiculatus* genomic DNA and salivary gland transcripts

Domains within proteins, which enable transposition of retroposons, such as reverse transcriptases, endonucleases and integrases, were used to search RaGI and BAC

© University of Pretoria

sequences using TblastX. Amongst the significant matches was a putative long interspersed element (LINE) (GB Accession: EU018129). LINEs, an acronym for Long Terminal Repeat (LTR) retrotransposons represent the second most prevalent subclass of transposable elements in vertebrate genomes, that uses a mechanism called target-primed reverse transcription (TPRT) for retrotransposition and integration into the genome. The LINE transposons contains two ORFs – ORF1 encodes a protein that binds to DNA, and ORF2, a bi-functional protein that encodes a domain of a reverse transcriptase and an endonuclease. The ORFs are separated by intergenic sequences that are flanked by UTRs. The 5' UTR has promoter activity while the 3' UTR is adjacent to a polyA tail. The putative LINE present in EU018129 is approximately 12.8 Kb in length and contains protein domains with sequence similarity to reverse transcriptases and endonucleases.

6.2.2.1 R2 LINE element in RaGI

Sequence fragments derived from several RA ESTs provide evidence for the presence of R2 elements in RaGI. R2, a LINE encoding a ~1,000 aa ORF, has been found to insert frequently into the gene encoding large subunit (LSU) rRNA in arthropods (Bunikis and Barbour, 2005). Features of the R2 element described in previous studies (Yang et al. 1999; Burke et al. 1999) are represented diagrammatically in Figure 6.2. R2 contains one ORF, which can be divided into three domains that exhibit similarity to functional domains in a range of proteins. A reverse transcriptase domain is positioned in the centre of the ORF. The N-terminal domain contains two motifs – a zinc-finger CCHH-like motif and a c-myb motif. The C-terminal domain contains an endonuclease. The ORF is flanked by untranslated regions (UTRs), which are responsible for the length variation observed in R2 elements. The 3' UTR is sometimes attached to a polyA tail (Bunikis and Barbour, 2005).

© University of Pretoria

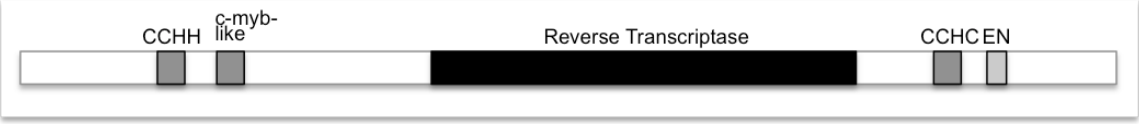


Figure 6.2 Diagrammatic representation of functional motifs present in the ORF of R2 retroelement. Black shaded box indicates the position of the reverse transcriptase domain. Dark grey boxes represent DNA-binding zinc-finger motifs CCHH and CCHC, and c-myb-like motif. Light grey box represents the endonuclease motif.

One EST in RaGI (CD782273), 679 bp in length, that has 99% sequence identity with 28S rRNA over 224 bp from its 3'end contains a poly-A sequence immediately upstream of this region (Figure 6.3). A 12 bp retrotransposon insertion site sequence 5'-TAGCCAAATGCC-3' that is conserved in other arthropods is also present downstream of the poly-A region.

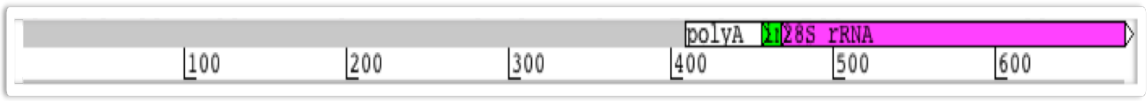


Figure 6.3 Diagrammatic representation of features on CD782273 (grey bar). Pink bar represents the region aligning with 28S rRNA; green box marks the insertion site for R2 retroelement; white bar highlights the polyA region.

A ~400 bp sequence upstream of the polyA tract present within CD782273 has 98% nucleotide identity over 365 bp with a different transcript CD779511. The CD779511 sequence has no significant similarity (E-value <1e-10) with sequences in GenBank, at either nucleotide or protein levels. However, the CD779511 EST is derived from the 3' end of a transcript. The 5' end of this same cloned cDNA, CD779512, is a singleton that is 735 bp in length. BlastX searching of CD779512 against the GenBank NR database identified weak matches to proteins with Zinc-finger motifs present within several other organisms. Zinc-finger motifs are involved in DNA binding. The presence of this motif prompted me to take a closer look at the motifs present in CD779512. I observed the two DNA-binding motifs - Zinc-finger CCHH-type and c-myb – that are characteristically present at the 5' end of an R2 element (Figure 6.4). This finding

© University of Pretoria

strongly suggests that the ~400 bp region upstream of the polyA sequence in CD782273, as well as the homologous sequence present in CD779511 represents the 3' UTR domain of an *R. appendiculatus* R2 LINE.

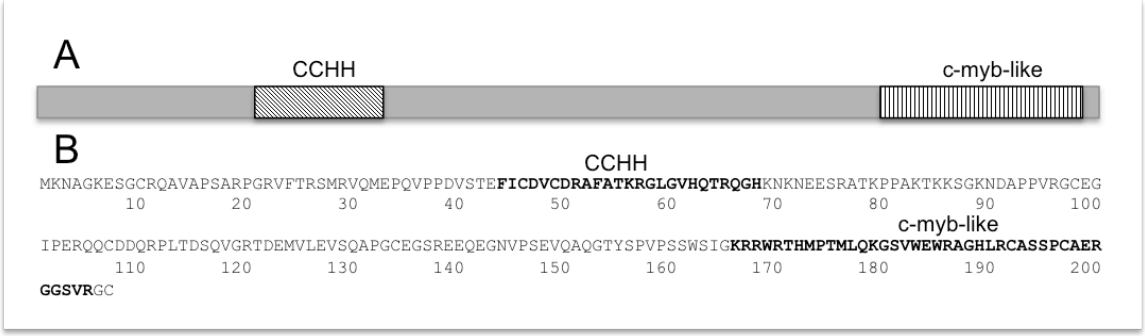


Figure 6.4 DNA-binding motifs of R2 retroelement located on CD779512 ORF. A. The grey bar represents 207 aa of the ORF; CCHH and c-myb-like motifs are represented by lined boxes. B. Amino acid sequence of the ORF with the two motifs represented in bold.

As previously mentioned, SINEs are non-autonomous transposable elements and therefore do not encode the machinery required for transcription and insertion into the genome. Movement of tRNA-derived SINEs is mediated by proteins encoded within LINEs. It is thought that for most categories of SINEs, each element has a specific LINE partner with which it shares similarity in the 3′ end sequence. The common 3′ ends are putative binding regions for the proteins that mediate retroposition (Okada et al., 1997). By contrast mammalian L1 and Alu SINE sequence families do not share 3′ homology with LINEs and their transposition may therefore be mechanistically distinct. The LINEs identified in *R. appendiculatus* do not appear to represent potential partners that might mediate transposition of Ruka since there is no detectable sequence similarity between their respective 3′ ends and Ruka. If the SINEs identified in the genomic contigs in this study are mobile, the genes required for their transposition do not appear to be encoded in closely adjacent regions of the genome. Direct comparison of DNA from different *R. appendiculatus* isolates would be required to determine whether the Ruka elements that I have discovered are actively mobile.

© University of Pretoria

6.2.2.2 Additional retrotransposons

Searches of the Censor database of repetitive elements (Kohany et al., 2006) strongly suggest that additional retroposons are present in the *R. appendiculatus* genome. Several LTRs and putative reverse transcriptase (RT) ORFs were identified in the *R. appendiculatus* BAC sequences by sequence similarity. Two putative Gypsy-like LTRs a class of retroposons described from *Drosophila* and four ORFs potentially encoding reverse transcriptase were identified in the BAC contig RAHE. Appendix Table A1.5 (a) summarizes the different classes of transposable elements to which the RA genomic sequences exhibited matches. In summary, over 500 sequences within RaGI had similarities to transposable elements, 26 were probably pseudogenes and more than 30 were similar to the PEN family of interspersed repeat sequences found in *Drosophila* (Haynes et al., 1987). PEN repeats are clusters of GGN repeats (where N is any nucleotide) that encode Glycine, resulting in Glycine-rich ORFs. The repeats are interspersed with a multiple of three nucleotides, which maintains in-frame protein translation within the repeat cluster. These repeats are commonly observes in long ORFs. The number of sequences for each category of repeat element is summarized in Appendix Table A1.5 (b). Forty-five sequences among those identified as repeat elements through searching the Censor database were analysed in more detail. Four sequences contained the ixodid SINE element Ruka which has already been described in depth; 18 contained at least one LTR domain, either reverse transcriptase, integrase or gag; seven had characteristics of a non-LTR reverse transcriptase; two appeared to represent candidate Class II transposable elements namely En/Spm described from Hymenoptera (wasps) and the Tc element known from the red flour beetle (*Tribolium castaneum*). Class II transposable elements that transpose by DNA duplication and insertion are frequently less abundant and more unevenly distributed within genomes than Class I retroposons (Kidwell, 2002; Wong and Choo, 2004). An exhaustive search for repeat elements was not performed. However, it is evident even from this preliminary analysis that a large proportion of RA transcribed sequences contain putative transposable elements.

© University of Pretoria

6.3 Conclusion

Given the abundance of transposable elements in the transcribed as well as genomic contigs of *R. appendiculatus*, in combination with the fact that no regions of the genomic contigs exhibited similarity over regions longer than 300 bp with the *I. scapularis* genomic contigs, the genome coverage of which is estimated to be > 80% (1.76 Gb) as given in Vectorbase assembly (IscaW1, 2008), it is evident that the genomes of ticks are rich in highly and moderately repetitive DNA and therefore structurally akin to those of vertebrates.

© University of Pretoria

Chapter 7. Concluding Remarks

This thesis describes manual curation and comparative analysis of EST databases and sample genomic sequences derived from the tick *Rhipicephalus appendiculatus*, to gain insight into the transcriptome and genomic organization of an ixodid tick which is an important vector of the protozoan parasite *Theileria parva*.

Nene et al. (2004) reported the sequencing of two cDNA libraries from 4-day fed uninfected and *Theileria parva*-infected *Rhipicephalus appendiculatus* ticks which derived 9,162 ESTs from the former and 9,844 from the latter library following RNA extraction and size selection of RNA that was > 2 kb in length. They reported no significant differential expression in the abundantly expressed ESTs between the two libraries, except for certain glycine-rich proteins, which were up-regulated in the infected library. However, a later study undertaken at ILRI identified several differentially expressed genes using the Suppression Subtractive Hybridization method (unpublished), some of which were present in RaGI. A majority of the differentially expressed genes from the SSH experiment did not have significant sequence similarity to any proteins in the GenBank database. However, one of these genes was up-regulated by 300% in *T. parva* infected, relative to uninfected, salivary glands. Structural analyses of this protein using threading demonstrated that the very short 69 aa ORF exhibited a fold similar to that of disintegrins - inhibitors of platelet aggregation. Similar threading approaches revealed the presence of conserved folds suggesting that additional unannotated transcripts may contain complement component and T-cell receptor homologues.

We sequenced additional cDNA clones derived from the mRNAs of uninfected salivary glands. This cDNA library was derived from identical RNA but was not size selected. This analysis resulted in the identification 1,784 additional ESTs, of which 1,202 (approximately 60%) that were novel. The mean length of ESTs sequenced without size-selection was significantly shorter than that of the size-selected dataset. The data

© University of Pretoria

indicates that technical biases in production can significantly affect the composition of EST datasets.

Sequencing of ~14,000 clones, without normalization was not close to providing complete coverage of the *R. appendiculatus* salivary gland transcriptome, as is evident by the absence of previously characterized proteins. Size selection also had a measurable effect on the number of short transcripts represented in the cDNA library and therefore risks overlooking transcripts of functional importance. Although obtaining an exhaustive catalogue of genes transcribed in any species may not feasible using the technology currently available. The new high throughput sequencing technologies, which are more cost effective, will provide much higher coverage than can be obtained by EST sequencing and come closer to achieving this goal.

Around 60% of transcripts in RaGI had no sequence homology to a known protein however a proportion were conserved in other tick species implying they could be of functional importance. Putative noncoding RNA sequence, in some cases present in multicopy families were also identified in RaGI.

There were 164,136,731 sequences amounting to 151,178,979,155 bases in GenBank (Release 195.0; April 15 2013). Of these, 6,816,360 sequences, making up 14,360,687 nucleotides originated from arthropods. That amounts to 4.15% of sequences and 0.0094% of the total nucleotides in GenBank. It is therefore not surprising to find that ~60% of transcripts in RaGI were not assigned a function using sequence homology-based methods. Using domain profile and structure-based methods I was able to assign putative annotations to some previously unannotated sequences however, more sophisticated algorithms have become available since these analyses were done, which should enable functional assignment for additional transcripts.

Tick genomes are comparable in size to those of vertebrates. This study estimated over 65,000 copies of a single family of mobile elements, specifically the SINE Ruka to be present in the *R. appendiculatus* genome. Over 500 copies of other sequences containing

© University of Pretoria

transposable element motifs were detected within RaGI and a relatively small random sample of genomic sequences thereby suggesting that the large genome size of ixodid ticks could at least partially be attributed to the presence of transposon-like sequences.

7.1 Future avenues for investigation

This study has focused on *in silico* analyses. The more significant discoveries should be validated using functional techniques, among which transcript knockdown, using RNAi has proved to be very powerful in many species including ixodid ticks.

High throughput analyses using the newer and cheaper sequencing platforms such as Illumina applied to multiple time-points and tissues are required to provide a more comprehensive analysis of the ixodid transcriptome.

The largest tick EST and genomic sequence dataset currently available for any tick is derived from the *I. scapularis* genome project. However this appears to have been subject to very limited comparative analysis with other species to date. Such a study would add considerable value to the field of arthropod genomics.

The apparent analogy between the organization of tick and vertebrate genomes is interesting, but does not extend to all arthropods, for example the Drosophila genome is relatively compact. It would be interesting to investigate the factors underlying these differences.

© University of Pretoria

Bibliography

Adler, M., Lazarus, R. A., Dennis, M. S. and Wagner, G. (1991). Solution structure of kistrin, a potent platelet aggregation inhibitor and GP IIb-IIIa antagonist. *Science* **253**, 445.

Akhmanova, A. S., Bindels, P. C., Xu, J., Miedema, K., Kremer, H. and Hennig, W. (1995). Structure and expression of histone H3.3 genes in Drosophila melanogaster and Drosophila hydei. *Genome* **38**, 586–600.

Alarcon-Chaidez, F. J., Sun, J. and Wikel, S. K. (2007). Transcriptome analysis of the salivary glands of Dermacentor andersoni Stiles (Acari: Ixodidae). *Insect biochemistry and molecular biology* **37**, 48–71.

Aljamali, M., Hern, L., Kupfer, D., Downard, S., So, S., Roe, B., Sauer, J. R. and Essenberg, R. (2009). Transcriptome analysis of the salivary glands of the female tick Amblyomma americanum. *Insect Molecular Biology* **18**, 129–154.

Almazán, C., Kocan, K. M., Garcia-garcia, J. C., Bergman, D. K., Garcia-Garcia, J. C., Blouin, E. F. and de La Fuente, J. (2003). Identification of protective antigens for the control of Ixodes scapularis infestations using cDNA expression library immunization. *Vaccine* **21**, 1492–501.

Anatriello, E., Ribeiro, J. M. C., de Miranda-Santos, I. K. F., Brandão, L. G., Anderson, J. M., Valenzuela, J. G., Maruyama, S. R., Silva, J. S. and Ferreira, B. R. (2010). An insight into the sialotranscriptome of the brown dog tick, Rhipicephalus sanguineus. *BMC genomics* **11**, 450.

Anderson, J. F. (2002). Natural History of Ticks. *Medical Clinics of North America* **86**, 205–218.

Andreotti, R., Gomes, A., Malavazi-Piza, K. C., Sasaki, S. D., Sampaio, C. A. M. and Tanaka, A. S. (2002). BmTI antigens induce a bovine protective immune response against Boophilus microplus tick. *International immunopharmacology* **2**, 557–63.

Andrews, N. W. and Webster, P. (1991). Phagolysosomal escape by intracellular pathogens. *Parasitology today (Personal ed.)* **7**, 335–40.

Arcà, B., Lombardo, F., Valenzuela, J. G., Francischetti, I. M. B., Marinotti, O., Coluzzi, M. and Ribeiro, J. M. C. (2005). An updated catalogue of salivary gland transcripts in the adult female mosquito, Anopheles gambiae. *The Journal of experimental biology* **208**, 3971–86.

Arrial, R. T., Togawa, R. C. and Brigido, M. D. M. (2009). Screening non-coding RNAs in transcriptomes from neglected species using PORTRAIT : case study of the pathogenic fungus Paracoccidioides brasiliensis. *BMC Bioinformatics* **9**, 1–9.

Bailey, T. L. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings International Conference on Intelligent Systems for Molecular Biology* **2**, 28–36.

Baldwin, C. L., Black, S. J., Brown, W. C., Conrad, P. A., Goddeeris, B. M., Kinuthia, S. W., Lalor, P. A., MacHugh, N. D., Morrison, W. I. and Morzaria, S. P. (1988). Bovine T cells, B cells, and null cells are transformed by the protozoan parasite Theileria parva. *Infection and immunity* **56**, 462–7.

Barker, S. C. and Murrell, A. (2004). Systematics and evolution of ticks with a list of valid genus and species names. *Parasitology* **129**, S15–S36.

© University of Pretoria

Appendix 1. Table A1.1

Barry, S. T., Ludbrook, S. B., Murrison, E. and Horgan, C. M. (2000). A regulated interaction between alpha5beta1 integrin and osteopontin. *Biochemical and biophysical research communications* **267**, 764–9.

Beckett, G. J. and Arthur, J. R. (2005). Selenium and endocrine systems. *The Journal of endocrinology* **184**, 455–65.

Beckmann, K., Grskovic, M., Gebauer, F. and Hentze, M. W. (2005). A dual inhibitory mechanism restricts msl-2 mRNA translation for dosage compensation in Drosophila. *Cell* **122**, 529–40.

Bergman, D. K., Palmer, M. J., Caimano, M. J., Radolf, J. D. and Wikel, S. K. (2000). Isolation and molecular cloning of a secreted immunosuppressant protein from Dermacentor andersoni salivary gland. *The Journal of parasitology* **86**, 516–25.

Bishop, R., Lambson, B., Wells, C., Pandit, P., Osaso, J., Nkonge, C., Morzaria, S., Musoke, A. and Nene, V. (2002). A cement protein of the tick Rhipicephalus appendiculatus, located in the secretory e cell granules of the type III salivary gland acini, induces strong antibody responses in cattle. *International journal for parasitology* **32**, 833–842.

Bishop, R. P., Odongo, D. O., Mann, D. J., Pearson, T., Sugimoto, C., Haines, L., Glass, E., Jensen, K., Seitzer, U., Ahmed, J. S., et al. (2009). Theileria. In *Genome Mapping and Genomics in Animal-Associated Microbes* (ed. Nene, V. M. and Kole, C.), pp. 191–231. Springer-Verlag Berlin Heidelberg.

Bolívar, J., Díaz, I., Iglesias, C. and Valdivia, M. M. (1999). Molecular cloning of a zinc finger autoantigen transiently associated with interphase nucleolus and mitotic centromeres and midbodies. Orthologous proteins with nine CXXC motifs highly conserved from nematodes to humans. *The Journal of biological chemistry* **274**, 36456–64.

Bowman, A. S. and Sauer, J. R. (2005). Tick salivary glands: function, physiology and future. *Parasitology* **129**, S67.

Bunikis, J. and Barbour, A. G. (2005). Ticks have R2 retrotransposons but not the consensus transposon target site of other arthropods. *Insect molecular biology* **14**, 465–74.

Chmelař, J., Anderson, J. M., Mu, J., Jochim, R. C., Valenzuela, J. G. and Kopecký, J. (2008). Insight into the sialome of the castor bean tick, Ixodes ricinus. *BMC Genomics* **9**, 1–21.

Claverie, J. (2005). Fewer genes, more noncoding RNA. *Science (New York, N.Y.)* **309**, 1529–30.

Clifford, R. J. and Kaplan, J. H. (2009). Regulation of Na,K-ATPase subunit abundance by translational repression. *The Journal of biological chemistry* **284**, 22905–15.

Couvreur, B., Beaufays, J., Charon, C., Lahaye, K., Gensale, F., Denis, V., Charloteaux, B., Decrem, Y., Prévôt, P.-P., Brossard, M., et al. (2008). Variability and action mechanism of a family of anticomplement proteins in Ixodes ricinus. *PloS one* **3**, e1400.

Crosnier, C., Bustamante, L. Y., Bartholdson, S. J., Bei, A. K., Theron, M., Uchikawa, M., Mboup, S., Ndir, O., Kwiatkowski, D. P., Duraisingh, M. T., et al. (2011). Basigin is a receptor essential for erythrocyte invasion by Plasmodium falciparum. *Nature* **480**, 534–7.

de Castro, J. J., James, A. D., Minjauw, B., Di Giulio, G. U., Permin, A., Pegram, R. G., Chizyuka, H. G. and Sinyangwe, P. (1997). Long-term studies on the economic impact of ticks on Sanga cattle in Zambia. *Experimental & applied acarology* **21**, 3–19.

© University of Pretoria

de Vos, S., Zeinstra, L., Taoufik, O., Willadsen, P. and Jongejan, F. (2001). Evidence for the utility of the Bm86 antigen from Boophilus microplus in vaccination against other tick species. *Experimental & applied acarology* **25**, 245–61.

de la Fuente, J. and Kocan, K. M. (2006). Strategies for development of vaccines for control of ixodid tick species. *Parasite immunology* **28**, 275–83.

de la Fuente, J., Rodríguez, M., Redondo, M., Montero, C., García-García, J. C., Méndez, L., Serrano, E., Valdés, M., Enriquez, A., Canales, M., et al. (1998). Field studies and cost-effectiveness analysis of vaccination with Gavac against the cattle tick Boophilus microplus. *Vaccine* **16**, 366–73.

de la Fuente, J., Almazán, C., Blas-Machado, U., Naranjo, V., Mangold, A. J., Blouin, E. F., Gortazar, C. and Kocan, K. M. (2006a). The tick protective antigen, 4D8, is a conserved protein involved in modulation of tick blood ingestion and reproduction. *Vaccine* **24**, 4082–95.

de la Fuente, J., Almazán, C., Naranjo, V., Blouin, E. F. and Kocan, K. M. (2006b). Synergistic effect of silencing the expression of tick protective antigens 4D8 and Rs86 in Rhipicephalus sanguineus by RNA interference. *Parasitology research* **99**, 108–13.

de la Fuente, J., Canales, M. and Kocan, K. M. (2006c). The importance of protein glycosylation in development of novel tick vaccine strategies. *Parasite immunology* **28**, 687–8.

Desjardins, C. A., Gundersen-rindal, D. E., Hostetler, J. B., Tallon, L. J., Fuester, R. W., Schatz, M. C., Pedroni, M. J., Fadrosh, D. W., Haas, B. J., Toms, B. S., et al. (2007). Structure and evolution of a proviral locus of Glyptapanteles indiensis bracovirus. *BMC Microbiology* **17**, 1–17.

Di Giulio, G., Lynen, G., Morzaria, S., Oura, C. and Bishop, R. (2009). Live immunization against East Coast fever--current status. *Trends in parasitology* **25**, 85–92.

Dinger, M. E., Pang, K. C., Mercer, T. R., Mattick, J. S., Pang, K. C. and Mercer, T. R. (2008). Differentiating Protein-Coding and Noncoding RNA: Challenges and Ambiguities. *PLoS Computational Biology* **4**,.

Dodgson, J. B., Yamamoto, M. and Engel, J. D. (1987). Chicken histone H3.3B cDNA sequence confirms unusual 3' UTR structure. *Nucleic acids research* **15**, 6294.

Düvel, K., Pries, R. and Braus, G. H. (2003). Polyadenylation of rRNA- and tRNA-based yeast transcripts cleaved by internal ribozyme activity. *Current genetics* **43**, 255–62.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Matrix* **32**, 1792–1797.

Fang, K. S., Vitale, M., Fehlner, P. and King, T. P. (1988). cDNA cloning and primary structure of a white-face hornet venom allergen, antigen 5. *Proceedings of the National Academy of Sciences of the United States of America* **85**, 895–9.

Francischetti, I. M. B., Valenzuela, J. G., Pham, V. M., Garfield, M. K. and Ribeiro, J. M. C. (2002). Toward a catalog for the transcripts and proteins (sialome) from the salivary gland of the malaria vector Anopheles gambiae. *The Journal of experimental biology* **205**, 2429–51.

Francischetti, I. M. B., Mather, T. N. and Ribeiro, J. M. C. (2003). Cloning of a salivary gland metalloprotease and characterization of gelatinase and fibrin(ogen)lytic activities in the saliva of the Lyme disease tick vector Ixodes scapularis. *Biochemical and biophysical research communications* **305**, 869–75.

© University of Pretoria

Appendix 1. Table A1.1

Francischetti, I. M. B., Mans, B. J., Meng, Z., Gudderra, N., Veenstra, T. D., Pham, V. M. and Ribeiro, J. M. C. (2008a). An insight into the sialome of the soft tick, Ornithodorus parkeri. *Insect biochemistry and molecular biology* **38**, 1–21.

Francischetti, I. M. B., Meng, Z., Mans, B. J., Gudderra, N., Hall, M., Veenstra, T. D., Pham, V. M., Kotsyfakis, M. and Ribeiro, J. M. C. (2008b). An insight into the salivary transcriptome and proteome of the soft tick and vector of epizootic bovine abortion, Ornithodoros coriaceus. *Journal of proteomics* **71**, 493–512.

Francischetti, I. M. B., Sa-Nunes, A., Mans, B. J., Santos, I. M. and Ribeiro, J. M. C. (2009). The role of saliva in tick feeding. *Frontiers in bioscience : a journal and virtual library* **14**, 2051–88.

Francischetti, I. M. B., Calvo, E., Andersen, J. F., Pham, V. M., Favreau, A. J., Barbian, K. D., Romero, A., Valenzuela, J. G. and Ribeiro, J. M. C. (2010). Insight into the Sialome of the Bed Bug, Cimex lectularius. *Journal of proteome research* **9**, 3820–31.

Francischetti, I. M. B., Anderson, J. M., Manoukis, N., Pham, V. M. and Ribeiro, J. M. C. (2011). An insight into the sialotranscriptome and proteome of the coarse bontlegged tick, Hyalomma marginatum rufipes. *Journal of proteomics* **74**, 2892–908.

Fujii, Y., Okuda, D., Fujimoto, Z., Horii, K., Morita, T. and Mizuno, H. (2003). Crystal structure of trimestatin, a disintegrin containing a cell adhesion recognition motif RGD. *Journal of molecular biology* **332**, 1115–22.

Galindo, M. I., Pueyo, J. I., Fouix, S., Bishop, S. A. and Couso, J. P. (2007). Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS biology* **5**, e106.

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

Godornes, C., Leader, B. T., Molini, B. J., Centurion-Lara, A. and Lukehart, S. A. (2007). Quantitation of rabbit cytokine mRNA by real-time RT-PCR. *Cytokine* **38**, 1–7.

Graf, J. F., Gogolewski, R., Leach-Bing, N., Sabatini, G. A., Molento, M. B., Bordin, E. L. and Arantes, G. J. (2004). Tick control: an industry point of view. *Parasitology* **129 Suppl**, S427–42.

Guerrero, F. D., Miller, R. J., Rousseau, M.-E., Sunkara, S., Quackenbush, J., Lee, Y. and Nene, V. M. (2005). BmiGI: a database of cDNAs expressed in Boophilus microplus, the tropical/southern cattle tick. *Insect biochemistry and molecular biology* **35**, 585–95.

Guerrero, F. D., Nene, V. M., George, J. E., Barker, S. C. and Willadsen, P. (2006). Sequencing a New Target Genome: The Boophilis microplus Genome Project. *Journale of Medical Entomology* **43**, 9–16.

Guerrero, F. D., Moolhuijzen, P. M., Peterson, D. G., Bidwell, S., Caler, E., Bellgard, M., Nene, V. M. and Djikeng, A. (2010). Reassociation kinetics-based approach for partial genome sequencing of the cattle tick, Rhipicephalus (Boophilus) microplus. *BMC genomics* **11**, 374.

Haapasalo, K., Vuopio, J., Syrjänen, J., Suvilehto, J., Massinen, S., Karppelin, M., Järvelä, I., Meri, S., Kere, J. and Jokiranta, T. S. (2012). Acquisition of complement factor H is important for pathogenesis of Streptococcus pyogenes infections: evidence from bacterial in vitro survival and human genetic association. *Journal of immunology (Baltimore, Md. : 1950)* **188**, 426–35.

Havilio, M., Levanon, E. Y., Lerman, G., Kupiec, M. and Eisenberg, E. (2005). Evidence for abundant transcription of non-coding regions in the Saccharomyces cerevisiae genome. *BMC genomics* **6**, 93.

© University of Pretoria

Haynes, S. R., Rebbert, M. L., Mozer, B. A., Forquignon, F. and Dawid, I. B. (1987). pen repeat sequences are GGN clusters and encode a glycine-rich domain in a Drosophila cDNA homologous to the rat helix destabilizing protein. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 1819–23.

He, H., Cai, L., Skogerbø, G., Deng, W., Liu, T., Zhu, X., Wang, Y., Jia, D., Zhang, Z., Tao, Y., et al. (2006). Profiling Caenorhabditis elegans non-coding RNA expression with a combined microarray. *Nucleic acids research* **34**, 2976–83.

Hill, C. A. and Wikel, S. K. (2005). The Ixodes scapularis genome project: an opportunity for advancing tick research. *Trends Parasitol.* **21**, 151–153.

Hoffman, D. R. (2006). Hymenoptera venom allergens. *Clinical reviews in allergy & immunology* **30**, 109–28.

Holm, L. and Rosenström, P. (2010). Dali server: conservation mapping in 3D. *Nucleic acids research* **38**, W545–9.

Horton, P. and Nakai, K. (1997). Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology* **5**, 147–52.

Huang, X. and Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. *Genome Research* **9**, 868–877.

Imamura, S., da Silva Vaz Junior, I., Sugino, M., Ohashi, K. and Onuma, M. (2005). A serine protease inhibitor (serpin) from Haemaphysalis longicornis as an anti-tick vaccine. *Vaccine* **23**, 1301–11.

Imamura, S., Namangala, B., Tajima, T., Tembo, M. E., Yasuda, J., Ohashi, K. and Onuma, M. (2006). Two serine protease inhibitors (serpins) that induce a bovine protective immune response against Rhipicephalus appendiculatus ticks. *Vaccine* **24**, 2230–7.

Jasrapuria, S., Arakane, Y., Osman, G., Kramer, K. J., Beeman, R. W. and Muthukrishnan, S. (2010). Genes encoding proteins with peritrophin A-type chitin-binding domains in Tribolium castaneum are grouped into three distinct families based on phylogeny, expression and function. *Insect biochemistry and molecular biology* **40**, 214–27.

Jones, D. T. (1999a). Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. 195–202.

Jones, D. (1999b). GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *Journal of molecular biology* **287**, 797–815.

Kapranov, P., Willingham, A. T. and Gingeras, T. R. (2007a). Genome-wide transcription and the implications for genomic organization. *Nature reviews. Genetics* **8**, 413–23.

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L., et al. (2007b). RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science (New York, N.Y.)* **316**, 1484–8.

Kato, H., Anderson, J. M., Kamhawi, S., Oliveira, F., Lawyer, P. G., Pham, V. M., Sangare, C. S., Samake, S., Sissoko, I., Garfield, M., et al. (2006). High degree of conservancy among secreted salivary gland proteins from two geographically distant Phlebotomus duboscqi sandflies populations (Mali and Kenya). *BMC genomics* **7**, 226.

© University of Pretoria

Appendix 1. Table A1.1

Kemp, D., Stone, B. and Binnington, K. (1982). Tick attachment and feeding: role of the mouthparts, feeding apparatus, salivary gland secretions and the host response. In *Physiology of Ticks* (ed. Obenchain, F. D. and Galun, R.), pp. 119–168. Pergamon Press.

Kidwell, M. G. (2002). Transposable elements and the evolution of genome size in eukaryotes. *Genetica* **115**, 49–63.

Kohany, O., Gentles, A. J., Hankus, L. and Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC bioinformatics* **7**, 474.

Kohtz, J. D., Mady, R., Bi, C., Shah, P., Clark, B. S. and Feng, J. (2006). The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes & Development* **20**, 1470–1484.

Krezel, A. M., Wagner, G., Seymour-Ulmer, J. and Lazarus, R. A. (1994). Structure of the RGD protein decorsin: conserved motif and distinct function in leech proteins that affect blood clotting. *Science (New York, N.Y.)* **264**, 1944–7.

Labuda, M. and Nuttall, P. A. (2004). Tick-borne viruses. *Parasitology* **129**, S221 – S245.

Lai, R., Liu, H., Lee, W. H. and Zhang, Y. (2003). Two novel Bv8-like peptides from skin secretions of the toad Bombina maxima. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* **134**, 509–514.

Lambson, B., Nene, V. M., Obura, M., Shah, T., Pandit, P., Ole-Moiyoi, O., Delroux, K., Welburn, S., Skilton, R., de Villiers, E., et al. (2005). Identification of candidate sialome components expressed in ixodid tick salivary glands using secretion signal complementation in mammalian cells. *Insect molecular biology* **14**, 403–14.

Lawrie, C. H., Randolph, S. E. and Nuttall, P. A. (1999). Ixodes ticks: serum species sensitivity of anticomplement activity. *Experimental parasitology* **93**, 207–14.

Leboulle, G., Rochez, C., Louahed, J., Ruti, B., Brossard, M., Bollen, A. and Godfroid, E. (2002). Isolation of Ixodes ricinus salivary gland mRNA encoding factors induced during blood feeding. *The American journal of tropical medicine and hygiene* **66**, 225–33.

Lee, G., Chan, W., Hurle, M. R., DesJarlais, R. L., Watson, F., Sathe, G. M. and Wetzel, R. (1993). Strong inhibition of fibrinogen binding to platelet receptor α IIb β 3 by RGD sequences installed into a presentation scaffold. *"Protein Engineering, Design and Selection"* **6**, 745–754.

Lee, J. Y., Park, J. Y. and Tian, B. (2008). Identification of mRNA polyadenylation sites in genomes using cDNA sequences, expressed sequence tags, and Trace. *Methods in molecular biology (Clifton, N.J.)* **419**, 23–37.

Li, S., Kwon, J. and Aksoy, S. (2001a). Characterization of genes expressed in the salivary glands of the tsetse fly, Glossina morsitans morsitans. *Insect molecular biology* **10**, 69–76.

Li, M., Bullock, C. M., Knauer, D. J., Ehlert, F. J. and Zhou, Q. Y. (2001b). Identification of two prokineticin cDNAs: recombinant proteins potently contract gastrointestinal smooth muscle. *Molecular pharmacology* **59**, 692–8.

Li, J., Ribeiro, J. M. C. and Yan, G. (2010). Allelic gene structure variations in Anopheles gambiae mosquitoes. *PloS one* **5**, e10699.

© University of Pretoria

Lightfoot, C. J. and Norval, R. A. I. (1981). Tick problems in wildlife in Zimbabwe. 1. The effects of tick parasitism on wild ungulates. *South African Journal of Wildlife Research* **11**, 41–45.

Lobley, A., Sadowski, M. I. and Jones, D. T. (2009). pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. *Bioinformatics (Oxford, England)* **25**, 1761–7.

Lowe, T. M. and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research* **25**, 955–64.

Mans, B. J., Louw, A. I. and Neitz, A. W. H. (2002). Savignygrin, a Platelet Aggregation Inhibitor from the Soft Tick Ornithodoros savignyi, Presents the RGD Integrin Recognition Motif on the Kunitz-BPTI Fold. *Journal of Biological Chemistry* **277**, 21371–21378.

Mans, B. J., Andersen, J. F., Francischetti, I. M. B., Valenzuela, J. G., Schwan, T. G., Pham, V. M., Garfield, M. K., Hammer, C. H. and Ribeiro, J. M. C. (2008a). Comparative sialomics between hard and soft ticks: implications for the evolution of blood-feeding behavior. *Insect biochemistry and molecular biology* **38**, 42–58.

Mans, B. J., Andersen, J. F., Schwan, T. G. and Ribeiro, J. M. C. (2008b). Characterization of anti-hemostatic factors in the argasid, Argas monolakensis: implications for the evolution of blood-feeding in the soft tick family. *Insect biochemistry and molecular biology* **38**, 22–41.

Mans, B. J., Ribeiro, J. M. C. and Andersen, J. F. (2008c). Structure, function, and evolution of biogenic amine-binding proteins in soft ticks. *The Journal of biological chemistry* **283**, 18721–33.

Maris, C., Dominguez, C. and Allain, F. H.-T. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *The FEBS journal* **272**, 2118–31.

Maritz-Olivier, C., Stutzer, C., Jongejan, F., Neitz, A. W. H. and Gaspar, A. R. M. (2007). Tick anti-hemostatics: targets for future vaccines and therapeutics. *Trends in parasitology* **23**, 397–407.

Marsden, R. L., McGuffin, L. J. and Jones, D. T. (2002). Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein science : a publication of the Protein Society* **11**, 2814–24.

Mattick, J. S. (2001). Non-coding RNAs: the architects of eukaryotic complexity. *EMBO reports* **2**, 986–91.

Mattick, J. S. (2003). Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays : news and reviews in molecular, cellular and developmental biology* **25**, 930–9.

Mattick, J. S. (2004). The hidden genetic program of complex organisms. *Scientific American* **291**, 60–7.

Mattick, J. S. (2005). The functional genomics of noncoding RNA. *Science (New York, N.Y.)* **309**, 1527–8.

Mattick, J. S. and Makunin, I. V. (2006). Non-coding RNA. *Human molecular genetics* **15 Spec No**, R17–29.

Maya-Monteiro, C. M., Daffre, S., Logullo, C., Lara, F. A., Alves, E. W., Capurro, M. L., Zingali, R., Almeida, I. C. and Oliveira, P. L. (2000). HeLp, a heme lipoprotein from the hemolymph of the cattle tick, Boophilus microplus. *The Journal of biological chemistry* **275**, 36584–9.

© University of Pretoria

Appendix 1. Table A1.1

McKenna, R. V., Riding, G. A., Jarmey, J. M., Pearson, R. D. and Willadsen, P. (1998). Vaccination of cattle against the Boophilus microplus using a mucin-like membrane glycoprotein. *Parasite immunology* **20**, 325–36.

McSwain, J. L., Essenberg, R. C. and Sauer, J. R. (1982). Protein changes in the salivary glands of the female lone star tick, Amblyomma americanum, during feeding. *The Journal of parasitology* **68**, 100–6.

Minjauw, B. and Mcleod, A. (2003). *Tick-borne diseases and poverty*.

Misra, S., Crosby, M. A., Mungall, C. J., Campbell, K. S., Hradecky, P., Huang, Y., Millburn, G. H., Prochnik, S. E., Smith, C. D., Tupy, J. L., et al. (2002). Annotation of the Drosophila melanogaster euchromatic genome: a systematic review. *Genome biology* **3**, 1–22.

Modrek, B., Resch, A., Grasso, C. and Lee, C. (2001). Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic acids research* **29**, 2850–9.

Mulenga, A., Sugimoto, C., Sako, Y., Ohashi, K., Musoke, A., Shubash, M. and Onuma, M. (1999). Molecular Characterization of a Haemaphysalis longicornis Tick Salivary Gland-Associated 29-kilodalton Protein and Its Effect as a Vaccine against Tick Infestation in Rabbits. *Infection and immunity* **67**, 1652–1658.

Mulenga, A., Sugino, M., Nakajima, M., Sugimoto, C. and Onuma, M. (2001). Tick encoded serine proteinase inhibitors (serpins); Potential target antigens for tick vaccine development. *J. Vet. Med. Sci.* **63**, 1063–1069.

Mulenga, A., Tsuda, A., Onuma, M. and Sugimoto, C. (2003). Four serine proteinase inhibitors (serpin) from the brown ear tick, Rhiphicephalus appendiculatus; cDNA cloning and preliminary characterization. *Insect biochemistry and molecular biology* **33**, 267.

Mulenga, A., Khumthong, R. and Chalaire, K. C. (2009). Ixodes scapularis tick serine proteinase inhibitor (serpin) gene family; annotation and transcriptional analysis. *BMC genomics* **10**, 217.

Murrell, A. and Barker, S. C. (2003). Synonymy of Boophilus Curtice, 1891 with Rhipicephalus Koch, 1844 (Acari: Ixodidae). *Systematic parasitology* **56**, 169–72.

Murrell, A., Campbell, N. J. and Barker, S. C. (2001). A total-evidence phylogeny of ticks provides insights into the evolution of life cycles and biogeography. *Molecular phylogenetics and evolution* **21**, 244–58.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., et al. (2000). A whole-genome assembly of Drosophila. *Science (New York, N.Y.)* **287**, 2196–204.

Nagasaki, H., Arita, M., Nishizawa, T., Suwa, M. and Gotoh, O. (2005). Species-specific variation of alternative splicing and transcriptional initiation in six eukaryotes. *Gene* **364**, 53–62.

Nakajima, C., Silva Vaz Jr, I. D., Imamura, S., Konnai, S., Ohashi, K. and Onuma, M. (2005). Random Sequencing of cDNA Library Derived from Partially-Fed Adult Female Haemaphysalis longicornis Salivary Gland. *J. Vet. Med. Sci.* **67**, 1127–1131.

Narasimhan, S., Koski, R. A., Beaulieu, B., Anderson, J. F., Ramamoorthi, N., Kantor, F., Cappello, M. and Fikrig, E. (2002). A novel family of anticoagulants from the saliva of Ixodes scapularis. *Insect molecular biology* **11**, 641–50.

Neitz, W. O. (1953). Aureomycin in Theileria parva infection. *Nature* **171**, 34–5.

© University of Pretoria

Nene, V. M., Lee, D., Quackenbush, J. and Skilton, R. (2002). AvGI, an index of genes transcribed in the salivary glands of the ixodid tick Amblyomma variegatum. *International journal for* **32**, 1447–56.

Nene, V. M., Lee, D., Kang'a, S., Skilton, R., Shah, T., de Villiers, E., Mwaura, S., Taylor, D., Quackenbush, J. and Bishop, R. (2004). Genes transcribed in the salivary glands of female Rhipicephalus appendiculatus ticks infected with Theileria parva. *Insect biochemistry and molecular biology* **34**, 1117–28.

Nijhof, A. M., Taoufik, A., de la Fuente, J., Kocan, K. M., de Vries, E. and Jongejan, F. (2007). Gene silencing of the tick protective antigens, Bm86, Bm91 and subolesin, in the one-host tick Boophilus microplus by RNA interference. *International journal for parasitology* **37**, 653–62.

Nobile, M., Noceti, F., Prestipino, G. and Possani, L. D. (1996). Helothermine, a lizard venom toxin, inhibits calcium current in cerebellar granules. *Experimental brain research. Experimentelle Hirnforschung. Expérimentation cérébrale* **110**, 15–20.

Norval, R. A., Andrew, H. R. and Yunker, C. E. (1989). Pheromone-mediation of host-selection in bont ticks (Amblyomma hebraeum koch). *Science (New York, N.Y.)* **243**, 364–5.

Norval, R. A. I., Perry, B. D. and Young, A. S. (1992). *The Epidemiology of Theileriosis in Africa*. illustrate. Academic Press.

Nuttall, P. A. and Labuda, M. (2004). Tick–host interactions: saliva-activated transmission. *Parasitology* **129**, S177–S189.

Okada, N. (1991). SINEs. *Current opinion in genetics & development* **1**, 498–504.

Okada, N., Hamada, M., Ogiwara, I. and Ohshima, K. (1997). SINEs and LINEs share common 3' sequences: a review. *Gene* **205**, 229–43.

Paesen, G. C., Adams, P. L., Harlos, K., Nuttall, P. A. and Stuart, D. I. (1999). Tick histamine-binding proteins: isolation, cloning, and three-dimensional structure. *Molecular cell* **3**, 661–71.

Paesen, G. C., Adams, P. L., Nuttall, P. A. and Stuart, D. L. (2000). Tick histamine-binding proteins : lipocalins with a second binding cavity. *Biochimica et Biophysica Acta* **1482**, 92–101.

Pesole, G. (2000). UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. *Nucleic Acids Research* **28**, 193–196.

Prevot, P.-P., Couvreur, B., Denis, V., Brossard, M., Vanhamme, L. and Godfroid, E. (2007). Protective immunity against Ixodes ricinus induced by a salivary serpin. *Vaccine* **25**, 3284–92.

Quackenbush, J., Liang, F., Holt, I., Pertea, G. and Upton, J. (2000). The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic acids research* **28**, 141–5.

Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R. and White, J. (2001). The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research* **29**, 159.

Radley, D. E., Brown, C. G. D., Cunningham, M. P., Kimber, C. D., Musisi, F. L., Payne, R. C., Purnell, R. E., Stagg, S. M. and Young, A. S. (1975a). East coast fever: 3. Chemoprophylactic immunization of cattle using oxytetracycline and a combination of theilerial strains. *Veterinary Parasitology* **1**, 51–60.

© University of Pretoria

Appendix 1. Table A1.1

Radley, D. E., Young, A. S., Brown, C. G. D., Burridge, M. J., Cunningham, M. P., Musisi, F. L. and Purnell, R. E. (1975b). East coast fever: 2. Cross-immunity trials with a Kenya strain of Theileria lawrencei. *Veterinary Parasitology* **1**, 43–50.

Ribeiro, J. M. C. (1987). Ixodes dammini: salivary anti-complement activity. *Experimental parasitology* **64**, 347–53.

Ribeiro, J. M. C. (1989). Role of saliva in tick/host interactions. *Experimental & applied acarology* **7**, 15–20.

Ribeiro, J. M. C. and Francischetti, I. M. B. (2003). Role of arthropod saliva in blood feeding: sialome and post-sialome perspectives. *Annual review of entomology* **48**, 73–88.

Ribeiro, J. M. C., Alarcon-Chaidez, F. J., Francischetti, I. M. B., Mans, B. J., Mather, T. N., Valenzuela, J. G. and Wikel, S. K. (2006). An annotated catalog of salivary gland transcripts from Ixodes scapularis ticks. *Insect biochemistry and molecular biology* **36**, 111–29.

Ribeiro, J. M. C., Arcà, B., Lombardo, F., My, V., Calvo, E., Phan, V. M., Chandra, P. K. and Wikel, S. K. (2007). An annotated catalogue of salivary gland transcripts in the adult female mosquito, Aedes aegypti. *BMC genomics* **8**, 6.

Rudenko, N., Golovchenko, M. and Grubhoffer, L. (2007). Gene organization of a novel defensin of Ixodes ricinus: first annotation of an intron/exon structure in a hard tick defensin gene and first evidence of the occurrence of two isoforms of one member of the arthropod defensin family. *Insect molecular biology* **16**, 501–7.

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

Sadd, B. M., Kube, M., Klages, S., Reinhardt, R. and Schmid-Hempel, P. (2010). Analysis of a normalised expressed sequence tag (EST) library from a key pollinator, the bumblebee Bombus terrestris. *BMC genomics* **11**, 110.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989). *Molecular Cloning: A Laboratory Manual*. second. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Santos, I. K. F. de M., Valenzuela, J. G., Ribeiro, J. M. C., de Castro, M., Costa, J. N., Costa, A. M., da Silva, E. R., Neto, O. B. R., Rocha, C., Daffre, S., et al. (2004). Gene discovery in Boophilus microplus, the cattle tick: the transcriptomes of ovaries, salivary glands, and hemocytes. *Annals of the New York Academy of Sciences* **1026**, 242–6.

Schneider, M. C., Prosser, B. E., Caesar, J. J. E., Kugelberg, E., Li, S., Zhang, Q., Quoraishi, S., Lovett, J. E., Deane, J. E., Sim, R. B., et al. (2009). Neisseria meningitidis recruits factor H using protein mimicry of host carbohydrates. *Nature* **458**, 890–3.

Shao, Y. and Ismail-Beigi, F. (2004). Control of Na+-K+-ATPase beta 1-subunit expression: role of 3'-untranslated region. *American journal of physiology. Cell physiology* **286**, C580–5.

Shaw, M. K. (1997). The same but different: the biology of Theileria sporozoite entry into bovine cells. *International journal for parasitology* **27**, 457–74.

Shedlock, A. M. and Okada, N. (2000). SINE insertions: powerful tools for molecular systematics. *BioEssays : news and reviews in molecular, cellular and developmental biology* **22**, 148–60.

Slomovic, S., Laufer, D., Geiger, D. and Schuster, G. (2006). Polyadenylation of ribosomal RNA in human cells. *Nucleic acids research* **34**, 2966–75.

© University of Pretoria

Sonnhammer, E., Von Heijne, G., Krogh, A. and others (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. In *Proc Int Conf Intell Syst Mol Biol*, pp. 175–82.

Sugino, M., Imamura, S., Mulenga, A., Nakajima, M., Tsuda, A., Ohashi, K. and Onuma, M. (2003). A serine proteinase inhibitor (serpin) from ixodid tick Haemaphysalis longicornis; cloning and preliminary assessment of its suitability as a candidate for a tick vaccine. *Vaccine* **21**, 2844–2851.

Sunter, J. D., Patel, S. P., Skilton, R. A., Githaka, N., Knowles, D. P., Scoles, G. A., Nene, V. M., de Villiers, E. and Bishop, R. P. (2008). A novel SINE family occurs frequently in both genomic DNA and transcribed sequences in ixodid ticks of the arthropod sub-phylum Chelicerata. *Gene* **415**, 13–22.

Sutton, G. G., White, O., Adams, M. D. and Kerlavage, A. R. (1995). TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*. **1**, 9–19.

Szeto, T. H., Wang, X. H., Smith, R., Connor, M., Christie, M. J., Nicholson, G. M. and King, G. F. (2000). Isolation of a funnel-web spider polypeptide with homology to mamba intestinal toxin 1 and the embryonic head inducer Dickkopf-1. *Toxicon : official journal of the International Society on Toxinology* **38**, 429–42.

Söderhäll, I., Kim, Y.-A., Jiravanichpaisal, P., Lee, S.-Y. and Söderhäll, K. (2005). An ancient role for a prokineticin domain in invertebrate hematopoiesis. *Journal of immunology (Baltimore, Md. : 1950)* **174**, 6153–60.

Tan, J., Liu, Z., Nomura, Y. and Goldin, A. (2002). Alternative splicing of an insect sodium channel gene generates pharmacologically distinct sodium channels. *The Journal of Neuroscience* **22**, 5300–5309.

Towle, D. W., Paulsen, R. S., Weihrauch, D., Kordylewski, M., Salvador, C., Lignot, J. H. and Spanings-Pierrot, C. (2001). Na(+)+K(+)-ATPase in gills of the blue crab Callinectes sapidus: cDNA sequencing and salinity-related expression of alpha-subunit mRNA and protein. *The Journal of experimental biology* **204**, 4005–12.

Trimnell, A. R., Hails, R. S. and Nuttall, P. A. (2002). Dual action ectoparasite vaccine targeting “exposed” and “concealed” antigens. *Vaccine* **20**, 3560–3568.

Uilenberg, G. (1999). Immunization against diseases caused by Theileria parva : a. *Tropical Medicine and International Health* **4**,.

Ullmann, A. J., Lima, C. M. R., Guerrero, F. D., Piesman, J. and Black, W. C. (2005). Genome size and organization in the blacklegged tick, Ixodes scapularis and the Southern cattle tick, Boophilus microplus. *Insect molecular biology* **14**, 217–22.

Valenzuela, J. G. (2004). Exploring tick saliva: from biochemistry to “sialomes” and functional genomics. *Parasitology* **129**, S83–S94.

Valenzuela, J. G., Charlab, R., Mather, T. N. and Ribeiro, J. M. C. (2000). Purification, cloning, and expression of a novel salivary anticomplement protein from the tick, Ixodes scapularis. *The Journal of biological chemistry* **275**, 18717–23.

Valenzuela, J. G., Francischetti, I. M. B., Pham, V. M., Garfield, M. K., Mather, T. N. and Ribeiro, J. M. C. (2002a). Exploring the sialome of the tick Ixodes scapularis. *Journal of Experimental Biology*. **205**, 2843–2864.

© University of Pretoria

Appendix 1. Table A1.1

Valenzuela, J. G., Pham, V. M., Garfield, M. K., Francischetti, I. M. B. and Ribeiro, J. M. C. (2002b). Toward a description of the sialome of the adult female mosquito Aedes aegypti. *Insect biochemistry and molecular biology* **32**, 1101–22.

van den Elsen, J. M. H., Martin, A., Wong, V., Clemenza, L., Rose, D. R. and Isenman, D. E. (2002). X-ray crystal structure of the C4d fragment of human complement component C4. *Journal of molecular biology* **322**, 1103–15.

Waladde, S. M. and Rice, M. (1982). The sensory basis of tick feeding behaviour. In *Physiology of Ticks* (ed. Obenchain, F. D. and Galun, R.), pp. 71–118. Pergamon Press, Oxford.

Wang, X., Coons, L. B., Taylor, D. B., Stevens, S. E. and Gartner, T. K. (1996). Variabilin, a Novel RGD-containing Antagonist of Glycoprotein IIb-IIIa and Platelet Aggregation Inhibitor from the Hard Tick Dermacentor variabilis. *The Journal of Biological Chemistry* **271**, 17785–17790.

Wang, H., Paesen, G. C., Nuttall, P. A. and Barbour, A. G. (1998). Male ticks help their mates to feed. *Nature* **391**, 753–754.

Wang, H., Kaufman, W. R., Cui, W. W. and Nuttall, P. A. (2001). Molecular individuality and adaptation of the tick Rhipicephalus appendiculatus in changed feeding environments. *Medical and veterinary entomology* **15**, 403–12.

Wang, M., Guerrero, F. D., Pertea, G. and Nene, V. M. (2007). Global comparative analysis of ESTs from the southern cattle tick, Rhipicephalus (Boophilus) microplus. *BMC genomics* **8**, 368.

Wang, P., Yu, P., Gao, P., Shi, T. and Ma, D. (2009). Discovery of novel human transcript variants by analysis of intronic single-block EST with polyadenylation site. *BMC genomics* **10**, 518.

Waxman, L., Smith, D. E., Arcuri, K. E. and Vlasuk, G. P. (1990). Tick anticoagulant peptide (TAP) is a novel inhibitor of blood coagulation factor Xa. *Science (New York, N.Y.)* **248**, 593–6.

Wikel, S. K. and Allen, J. R. (1978). Acquired resistance to ticks. III. Cobra venom factor and the resistance response. *Immunology* **32**, 457–65.

Willadsen, P. (2004). Anti-tick vaccines. *Parasitology* **129**, 367–387.

Willadsen, P., McKenna, R. and Riding, G. (1988). Isolation from the cattle tick, Boophilus microplus, of antigenic material capable of eliciting a protective immunological response in the bovine host. *International Journal for …* **18**, 183–9.

Willadsen, P., Riding, G. A., McKenna, R. V., Kemp, D. H., Tellam, R. L., Nielsen, J. N., Lahnstein, J., Cobon, G. S. and Gough, J. M. (1989). Immunologic control of a parasitic arthropod. Identification of a protective antigen from Boophilus microplus. *Journal of immunology (Baltimore, Md. : 1950)* **143**, 1346–51.

Willadsen, P., Bird, P., Cobon, G. S. and Hungerford, J. (1995). Commercialisation of a recombinant vaccine against Boophilus microplus. *Parasitology* **110 Suppl**, S43–50.

Willadsen, P., Smith, D., Cobon, G. and McKenna, R. V. (1996). Comparative vaccination of cattle against Boophilus microplus with recombinant antigen Bm86 alone or in combination with recombinant Bm91. *Parasite immunology* **18**, 241–6.

Wong, L. H. and Choo, K. H. A. (2004). Evolutionary dynamics of transposable elements at the centromere. *Trends in genetics : TIG* **20**, 611–6.

© University of Pretoria

Wu, T. D. and Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics (Oxford, England)* **21**, 1859–75.

Yao, A., Charlab, R. and Li, P. (2006). Systematic identification of pseudogenes through whole genome expression evidence profiling. *Nucleic acids research* **34**, 4477.

Young, A. S. and Leitch, B. L. (1980). A probable relationship between the development of Theileria species and the ecdysis of their tick hosts. *The Journal of parasitology* **66**, 356–9.

Zhang, Y. (2008). I-TASSER server for protein 3D structure prediction. *BMC bioinformatics* **9**, 40.

Zhu, S., Si, M.-L., Wu, H. and Mo, Y.-Y. (2007). MicroRNA-21 targets the tumor suppressor gene tropomyosin 1 (TPM1). *The Journal of biological chemistry* **282**, 14328–36.

© University of Pretoria

Appendix 1. Table A1.1

Appendix 1 Tables

Table A1.1 Best matches to known proteins in GenBank of 206 RaGI sequences that were unannotated in the initial study (Nene et al., 2004). Matches were obtained from BlastX sequence search against GenBank non-redundant protein database.

RaGI Identifier	Hit in GenBank	% identity	Alignment Length (aa)	E-value
TC1000	gi|241555295|ref|XP_002399426.1|	66.48	182	6.00E-35
TC1011	gi|242000258|ref|XP_002434772.1|	79.01	181	7.30E-69
TC1022	gi|260908481|gb|ACX53960.1|	87.86	140	1.80E-66
TC103	gi|241696184|ref|XP_002411837.1|	87.21	86	2.60E-33
TC1034	gi|241999414|ref|XP_002434350.1|	73.03	89	2.10E-27
TC1035	gi|241716962|ref|XP_002403885.1|	51.45	173	1.70E-43
TC1039	gi|241784935|ref|XP_002400484.1|	72.32	112	7.10E-37
TC1054	gi|241157424|ref|XP_002408042.1|	70.42	142	1.60E-45
TC1074	gi|241671866|ref|XP_002411435.1|	40.85	355	7.10E-71
TC1075	gi|241743214|ref|XP_002414190.1|	47.62	147	1.00E-28
TC1122	gi|242002832|ref|XP_002436059.1|	42.93	191	3.60E-35
TC1127	gi|241044323|ref|XP_002407185.1|	43	300	3.10E-62
TC1137	gi|291402903|ref|XP_002718246.1|	28.21	351	2.30E-36
TC1141	gi|241156201|ref|XP_002407719.1|	61.81	144	2.50E-34
TC1144	gi|241713546|ref|XP_002412099.1|	67.38	187	3.10E-67
TC1145	gi|241829811|ref|XP_002414779.1|	55	180	1.40E-45
TC1147	gi|241784742|ref|XP_002414407.1|	45.02	422	2.00E-103
TC115	gi|241705672|ref|XP_002403125.1|	42.42	422	1.80E-90
TC1183	gi|242000428|ref|XP_002434857.1|	59.5	121	9.30E-33
TC1184	gi|241680870|ref|XP_002411582.1|	78	150	1.60E-45
TC1206	gi|196476652|gb|ACG76192.1|	37.78	180	3.80E-33
TC1210	gi|241713103|ref|XP_002413477.1|	47.11	225	7.30E-54
TC1211	gi|241698619|ref|XP_002413132.1|	65.07	146	1.10E-42
TC1224	gi|241745186|ref|XP_002414256.1|	88.54	96	4.40E-45
TC1225	gi|241686712|ref|XP_002411689.1|	65.75	73	6.70E-30
TC1246	gi|241554591|ref|XP_002399519.1|	51.15	131	1.40E-45
TC1248	gi|242002908|ref|XP_002436097.1|	36.42	173	4.50E-26
TC1250	gi|260908312|gb|ACX53877.1|	85.59	111	2.10E-48
TC1254	gi|241268826|ref|XP_002406500.1|	27.96	372	4.10E-29
TC1255	gi|241841423|ref|XP_002415338.1|	70	110	1.30E-63
TC1309	gi|71726990|gb|AAZ39660.1|	57.1	345	1.00E-104
TC1310	gi|71726990|gb|AAZ39660.1|	59.46	296	6.00E-96
TC1312	gi|71726990|gb|AAZ39660.1|	56.86	255	1.50E-69
TC135	gi|67083573|gb|AAY66722.1|	68.81	109	7.10E-31
TC136	gi|67083573|gb|AAY66722.1|	68.81	109	7.60E-31
TC1381	gi|241173805|ref|XP_002410890.1|	65.66	166	5.90E-85
TC1413	gi|241333874|ref|XP_002408368.1|	64.49	459	6.00E-153
TC142	gi|260784045|ref|XP_002587080.1|	45.61	171	7.70E-33

© University of Pretoria

RaGI Identifier	Hit in GenBank	% identity	Alignment Length (aa)	E-value
TC1458	gi|241835842|ref|XP_002415070.1|	86.49	111	4.10E-45
TC1466	gi|45478108|gb|AAS66225.1|	63.22	87	6.20E-26
TC1469	gi|241693015|ref|XP_002412961.1|	81.98	111	2.30E-43
TC147	gi|242001660|ref|XP_002435473.1|	92.19	64	2.10E-28
TC1486	gi|241575249|ref|XP_002403179.1|	74.72	178	8.50E-52
TC1539	gi|260908312|gb|ACX53877.1|	82.88	111	3.90E-50
TC1543	gi|241153729|ref|XP_002407146.1|	44.44	207	1.30E-32
TC1548	gi|89277230|gb|ABD66751.1|	82.69	468	0
TC1551	gi|89954505|gb|ABD83654.1|	85.17	391	0
TC1557	gi|241655327|ref|XP_002411372.1|	75.65	230	2.70E-65
TC1565	gi|242002624|ref|XP_002435955.1|	48	125	1.00E-27
TC1578	gi|241837576|ref|XP_002415175.1|	83.89	149	9.70E-69
TC1581	gi|196476652|gb|ACG76192.1|	43.29	164	3.50E-44
TC1586	gi|241728328|ref|XP_002412248.1|	73.68	604	0
TC1589	gi|241029428|ref|XP_002406432.1|	59.14	279	2.60E-79
TC1593	gi|260908658|gb|ACX54048.1|	51.83	164	2.80E-48
TC1598	gi|241743784|ref|XP_002405409.1|	51.09	229	6.70E-86
TC1605	gi|241851527|ref|XP_002415777.1|	63.92	97	1.30E-28
TC1611	gi|241853825|ref|XP_002415932.1|	92.96	71	3.30E-30
TC1621	gi|241171432|ref|XP_002410646.1|	62.91	275	6.90E-89
TC1624	gi|33667952|gb|AAQ24553.1|	71.76	131	7.90E-49
TC1635	gi|241829407|ref|XP_002414760.1|	74.19	93	3.10E-35
TC1647	gi|241708457|ref|XP_002403278.1|	42.03	395	3.90E-65
TC1655	gi|196476652|gb|ACG76192.1|	39.76	166	4.60E-33
TC1663	gi|241999978|ref|XP_002434632.1|	52.07	217	2.30E-56
TC1674	gi|241063628|ref|XP_002408191.1|	52.74	146	1.20E-34
TC1693	gi|260908362|gb|ACX53902.1|	93.52	108	8.30E-46
TC1695	gi|241636946|ref|XP_002410667.1|	30.09	216	1.00E-27
TC1702	gi|89277230|gb|ABD66751.1|	62.69	469	9.00E-173
TC1714	gi|71726984|gb|AAZ39657.1|	48.12	478	2.00E-116
TC1741	gi|241595279|ref|XP_002404453.1|	54.68	278	2.50E-65
TC1742	gi|241846405|ref|XP_002415566.1|	68.85	183	3.80E-53
TC1757	gi|241690495|ref|XP_002411775.1|	80.88	136	4.80E-59
TC1764	gi|241561010|ref|XP_002401146.1|	48.08	287	2.10E-72
TC1777	gi|241999550|ref|XP_002434418.1|	69.3	228	5.90E-88
TC1785	gi|194241600|gb|ACF35055.1|	74.59	429	0
TC1799	gi|170285597|emb|CAO00628.1|	26.86	484	1.80E-37
TC1803	gi|242001436|ref|XP_002435361.1|	58.38	173	2.00E-39
TC1812	gi|241631831|ref|XP_002410291.1|	40.89	247	3.10E-47
TC1836	gi|241036101|ref|XP_002406797.1|	49.48	194	6.60E-46
TC1839	gi|152125831|gb|ABK40086.2|	68.45	317	1.00E-109
TC1846	gi|241836404|ref|XP_002415098.1|	50	136	1.70E-46
TC1868	gi|241716070|ref|XP_002413548.1|	45.41	207	5.50E-26
TC1875	gi|241028781|ref|XP_002406363.1|	46.92	260	1.60E-61
TC1898	gi|241618433|ref|XP_002408339.1|	52.91	189	3.20E-48
TC1907	gi|241852776|ref|XP_002415855.1|	47.1	155	7.00E-27
TC1916	gi|241997912|ref|XP_002433599.1|	81.18	85	1.80E-33

© University of Pretoria

Appendix 1. Table A1.1

RaGI Identifier	Hit in GenBank	% identity	Alignment Length (aa)	E-value
TC1928	gi|241733220|ref|XP_002412319.1|	50.55	182	3.20E-43
TC1936	gi|241388732|ref|XP_002409358.1|	35.62	292	2.90E-44
TC1952	gi|241602756|ref|XP_002405527.1|	98.06	155	2.80E-91
TC1984	gi|242001194|ref|XP_002435240.1|	70.59	85	3.20E-26
TC1988	gi|241594849|ref|XP_002404396.1|	56.47	340	7.50E-87
TC1994	gi|241756426|ref|XP_002401400.1|	78.91	147	1.60E-57
TC1996	gi|241603791|ref|XP_002405760.1|	86.62	314	2.00E-152
TC2012	gi|260908564|gb|ACX54001.1|	90.8	163	4.60E-78
TC2013	gi|241176114|ref|XP_002399487.1|	41.67	168	2.90E-31
TC2016	gi|241559453|ref|XP_002400802.1|	86.3	146	3.80E-50
TC2020	gi|241999932|ref|XP_002434609.1|	64.74	190	1.20E-47
TC2063	gi|51011418|gb|AAT92118.1|	65.26	190	1.10E-73
TC2064	gi|241786592|ref|XP_002414457.1|	39.12	501	1.10E-99
TC2069	gi|241779163|ref|XP_002399852.1|	88.57	70	2.50E-29
TC2076	gi|241593532|ref|XP_002404205.1|	70.25	121	5.50E-39
TC2082	gi|241653581|ref|XP_002410483.1|	59.06	127	4.70E-37
TC2089	gi|241104728|ref|XP_002409984.1|	67.95	78	2.80E-51
TC2120	gi|241680857|ref|XP_002411577.1|	73.68	171	6.20E-76
TC2127	gi|242002870|ref|XP_002436078.1|	48.79	414	5.00E-109
TC2132	gi|241997730|ref|XP_002433514.1|	64.36	289	2.30E-77
TC2137	gi|241680595|ref|XP_002412697.1|	52.74	146	4.50E-28
TC2146	gi|241997650|ref|XP_002433474.1|	70.89	158	9.90E-66
TC2151	gi|71027135|ref|XP_763211.1|	97.18	71	1.20E-32
TC2159	gi|241048575|ref|XP_002407299.1|	60.43	187	7.50E-47
TC2167	gi|241619905|ref|XP_002408610.1|	50.26	195	1.20E-40
TC2180	gi|71726990|gb|AAZ39660.1|	64.84	91	8.00E-39
TC2181	gi|241048565|ref|XP_002407296.1|	86.57	67	5.00E-31
TC2231	gi|241566790|ref|XP_002402186.1|	50.24	207	2.50E-52
TC2247	gi|241999558|ref|XP_002434422.1|	55.69	167	3.40E-54
TC2255	gi|241830508|ref|XP_002414805.1|	76.92	78	9.70E-29
TC2257	gi|241745698|ref|XP_002414272.1|	63.16	95	8.80E-27
TC2264	gi|240952158|ref|XP_002399330.1|	64.62	130	9.00E-37
TC2283	gi|241999830|ref|XP_002434558.1|	49.17	181	1.20E-40
TC2287	gi|241714822|ref|XP_002413529.1|	50	212	1.90E-47
TC2294	gi|260908610|gb|ACX54024.1|	43.27	245	3.10E-33
TC2302	gi|241617530|ref|XP_002406937.1|	49.24	329	2.10E-83
TC2307	gi|241151431|ref|XP_002406680.1|	47.32	224	8.50E-49
TC2324	gi|241682471|ref|XP_002411631.1|	73.97	242	1.70E-81
TC2330	gi|241600797|ref|XP_002404982.1|	83.87	62	2.20E-30
TC2334	gi|241114858|ref|XP_002400475.1|	57.89	152	8.90E-27
TC2341	gi|241696779|ref|XP_002413098.1|	53.88	245	4.00E-46
TC2364	gi|241757317|ref|XP_002401506.1|	73.17	41	8.80E-30
TC2376	gi|242000324|ref|XP_002434805.1|	67.66	167	6.80E-70
TC2434	gi|241603611|ref|XP_002405338.1|	63.39	183	1.90E-58
TC2446	gi|241834468|ref|XP_002414995.1|	49.74	193	6.20E-39
TC2451	gi|241171667|ref|XP_002410686.1|	80.95	84	4.60E-29
TC2476	gi|260908596|gb|ACX54017.1|	87.15	179	2.00E-81
TC2500	gi|242001820|ref|XP_002435553.1|	77.05	122	2.80E-48

© University of Pretoria

RaGI Identifier	Hit in GenBank	% identity	Alignment Length (aa)	E-value
TC2501	gi|241784599|ref|XP_002414402.1|	55.49	182	7.70E-57
TC2509	gi|241158243|ref|XP_002408276.1|	66.47	167	5.50E-61
TC2520	gi|241608423|ref|XP_002405980.1|	55.07	138	5.70E-32
TC2529	gi|241716070|ref|XP_002413548.1|	80.47	128	7.00E-55
TC253	gi|194241600|gb|ACF35055.1|	84.92	252	8.00E-129
TC254	gi|241685146|ref|XP_002412774.1|	64.17	120	1.30E-28
TC2543	gi|241240588|ref|XP_002401733.1|	50.94	159	3.60E-28
TC262	gi|241836404|ref|XP_002415098.1|	35.14	276	4.60E-47
TC273	gi|242024204|ref|XP_002432519.1|	56	100	2.70E-26
TC275	gi|241103946|ref|XP_002409944.1|	63.16	95	7.30E-27
TC278	gi|260908312|gb|ACX53877.1|	82.88	111	2.80E-51
TC288	gi|241723187|ref|XP_002413702.1|	46.56	189	9.00E-36
TC299	gi|241048974|ref|XP_002407315.1|	78.99	119	1.10E-49
TC308	gi|241747545|ref|XP_002414335.1|	77.27	88	4.60E-33
TC338	gi|242000540|ref|XP_002434913.1|	46.13	310	6.00E-40
TC35	gi|121308309|dbj|BAF43575.1|	46.72	274	5.60E-66
TC352	gi|118099923|ref|XP_420105.2|	33.33	138	6.70E-31
TC36	gi|67906164|dbj|BAE00066.1|	45.76	177	3.90E-38
TC37	gi|121308309|dbj|BAF43575.1|	39.82	221	2.00E-60
TC378	gi|241616954|ref|XP_002408092.1|	73.39	124	4.40E-30
TC410	gi|260908656|gb|ACX54047.1|	61.76	204	1.30E-65
TC419	gi|242002234|ref|XP_002435760.1|	50.87	173	1.00E-29
TC422	gi|241632497|ref|XP_002408615.1|	49.69	163	2.10E-35
TC443	gi|241853972|ref|XP_002415935.1|	62.29	236	6.70E-61
TC450	gi|260908420|gb|ACX53930.1|	98.72	78	3.30E-36
TC466	gi|241153731|ref|XP_002407147.1|	44.12	204	5.80E-36
TC480	gi|241112253|ref|XP_002399563.1|	50.21	468	2.00E-126
TC498	gi|240951961|ref|XP_002399283.1|	78.85	208	9.00E-93
TC503	gi|241557628|ref|XP_002400228.1|	46.93	179	5.00E-34
TC519	gi|241648009|ref|XP_002410005.1|	70.05	187	6.70E-76
TC549	gi|241730208|ref|XP_002413818.1|	67.42	132	2.00E-47
TC58	gi|67083415|gb|AAY66643.1|	69.39	98	5.00E-35
TC582	gi|241629285|ref|XP_002408268.1|	62.69	268	7.10E-91
TC588	gi|215500538|gb|EEC10032.1|	80.95	126	3.20E-81
TC606	gi|198414579|ref|XP_002130477.1|	29.66	236	5.00E-28
TC614	gi|241153729|ref|XP_002407146.1|	44	175	7.10E-32
TC630	gi|241054403|ref|XP_002407654.1|	87.5	144	2.50E-56
TC632	gi|241838712|ref|XP_002415211.1|	63.27	245	9.10E-88
TC636	gi|260908473|gb|ACX53956.1|	100	58	7.80E-28
TC643	gi|241741504|ref|XP_002412383.1|	71.79	78	8.50E-28
TC671	gi|241747129|ref|XP_002405614.1|	57.62	210	2.50E-67
TC677	gi|241166989|ref|XP_002409965.1|	78.76	113	3.40E-44
TC681	gi|241177505|ref|XP_002400066.1|	49.67	304	3.40E-53
TC729	gi|241568964|ref|XP_002402613.1|	65	100	1.40E-31
TC740	gi|241849741|ref|XP_002415701.1|	32.75	342	5.70E-38
TC75	gi|241779329|ref|XP_002399895.1|	58.09	136	1.80E-44
TC754	gi|241779900|ref|XP_002400040.1|	55.86	222	1.30E-63

© University of Pretoria

Appendix 1. Table A1.1

RaGI Identifier	Hit in GenBank	% identity	Alignment Length (aa)	E-value
TC758	gi|240973536|ref|XP_002401492.1|	75.13	193	8.60E-71
TC76	gi|241654403|ref|XP_002411323.1|	58.96	251	5.40E-81
TC761	gi|260908518|gb|ACX53978.1|	53.19	141	2.00E-35
TC762	gi|241743103|ref|XP_002414182.1|	76	75	1.10E-41
TC768	gi|241856245|ref|XP_002416055.1|	90.57	106	1.90E-43
TC789	gi|241999426|ref|XP_002434356.1|	80.61	98	9.00E-41
TC797	gi|241047150|ref|XP_002407223.1|	53.91	115	4.70E-28
TC804	gi|241735326|ref|XP_002413915.1|	65.94	138	6.20E-38
TC813	gi|298204323|gb|ADI61810.1|	45.11	133	2.80E-27
TC821	gi|241257861|ref|XP_002404675.1|	42.42	198	3.90E-35
TC828	gi|241722895|ref|XP_002404226.1|	40.87	208	2.20E-27
TC835	gi|241635882|ref|XP_002410577.1|	56.97	337	3.10E-85
TC847	gi|241680340|ref|XP_002412683.1|	54.93	142	8.30E-41
TC85	gi|242002670|ref|XP_002435978.1|	82.14	168	4.80E-70
TC856	gi|241250775|ref|XP_002403377.1|	53.14	175	1.90E-37
TC879	gi|241250775|ref|XP_002403377.1|	45.41	185	7.60E-35
TC888	gi|260836393|ref|XP_002613190.1|	35.16	219	2.00E-55
TC895	gi|241608964|ref|XP_002406064.1|	78.35	97	1.10E-34
TC904	gi|241613722|ref|XP_002407440.1|	61.72	303	8.00E-106
TC923	gi|149287044|gb|ABR23421.1|	82.19	73	1.40E-27
TC955	gi|292613042|ref|XP_002661721.1|	46.67	195	3.40E-42
TC956	gi|240994863|ref|XP_002404555.1|	61.02	118	1.00E-34
TC963	gi|241833190|ref|XP_002414930.1|	63.04	138	6.20E-50
TC965	gi|241836404|ref|XP_002415098.1|	41.21	165	7.20E-27
TC988	gi|240973536|ref|XP_002401492.1|	61.96	92	1.50E-27
TC989	gi|241997990|ref|XP_002433638.1|	50.76	132	6.10E-27

© University of Pretoria

Table A1.2 Sequences constituting gene families that are expressed in RaGI and their best match to GenBank's protein database (taken from Francischetti et al., 2009).

RaGI ID	Best match in GenBank[a]	Seq len.[c]	Signal Peptide present?	E-value	Len. of best match[b]	% identity
Glycine rich superfamily						
TC2529	cuticular protein (gi|74267392)-Tachypleus tridentatus	174	Yes	2.00E-14	212	49
CD783571	Cuticle protein 16.8 (Ir-ACP16.8) (gi|55976266)-	102	YES	4.00E-11	161	45
TC48	cuticular protein (gi|74267402)-Tachypleus tridentatus	156	YES	2.00E-40	159	54
TC48	conserved hypothetical protein (gi|108869047)-Aedes aegypti	482	No	1.00E-40	445	38
CD788108	hypothetical protein (gi|5911724)-Ixodes ricinus	196	YES	2.00E-45	252	54
TC2084	salivary gland protein (gi|50363178)-Ixodes scapularis	329	YES	6.00E-91	320	60
CD788013	salivary gland protein (gi|50363178)-Ixodes scapularis	227	YES	3.00E-53	320	54
TC1549	salivary gland protein (gi|50363178)-Ixodes scapularis	381	YES	3.00E-68	320	49
TC2184	hypothetical protein (gi|109238448)-Eimeria tenella	207	YES	5.00E-11	967	33
TC2256	hypothetical protein RRC32 [Uncultured methanogenic archaeon RC-I] (gi|147919123)-uncultured methanogenic archaeon RC-I	231	YES	2.00E-15	749	40
TC1421	Collagen alpha-2(IV) chain precursor gi|159649|gb|AAA18014.1| putative (gi|115347)-	550	No	0	1763	62
CD797106	PREDICTED: similar to Collagen alpha-5(IV) chain (gi|118095047)-Gallus gallus	267	YES	2.00E-57	1808	49
CD791446	type IV collagen alpha 5 chain (gi|27461187)-Canis familiaris	243	No	9.00E-65	1684	48
TC528	Collagen alpha-2(IV) chain precursor gi|159649|gb|AAA18014.1| putative (gi|115347)-	197	YES	4.00E-43	1763	61
TC180	Fibroin heavy chain precursor (Fib-H) (H-fibroin) (gi|9087216)-Bombyx mori	292	YES	1.00E-48	5263	41
TC180	Fibroin heavy chain precursor (Fib-H) (H-fibroin) (gi|9087216)-Bombyx mori	421	YES	2.00E-82	5263	46
CD780042	putative secreted salivary gland peptide (gi|56159959)-Ixodes scapularis	149	YES	7.00E-44	161	61
TC1134	salivary gland-associated protein 64P (gi|20069012)-Rhipicephalus appendiculatus	176	YES	1.00E-31	154	53
TC2381	salivary gland-associated protein 64P (gi|20069012)-Rhipicephalus appendiculatus	209	YES	7.00E-30	154	48
CD781439	ovarian fibroin-like substance-1 (gi|10954048)-Cyprinus carpio	227	YES	3.00E-19	421	44
TC1399	putative cement protein RIM36 (gi|21885262)-Rhipicephalus appendiculatus	311	YES	1.00E-177	334	97
TC1398	putative cement protein RIM36 (gi|21885262)-Rhipicephalus appendiculatus	334	YES	0	334	100

© University of Pretoria

Appendix 1. Table A1.2

RaGI ID	Best match in GenBank[a]	Seq len.[c]	Signal Peptide present?	E-value	Len. of best match[b]	% identity
TC1400	putative cement protein RIM36 (gi|21885262)-Rhipicephalus appendiculatus	338	YES	0	334	96
TC1003	Os03g0309300 [Oryza sativa (japonica cultivar-group)] (gi|115452621)-Oryza sativa (japonica cultivar-group)	184	No	1.00E-33	590	51
TC2265	hypothetical protein OsJ_023458 (gi|125600399)-Oryza sativa (japonica cultivar-group)	193	YES	9.00E-36	344	55
TC1955	glycine-rich protein (gi|7636182)-Triticum aestivum	323	YES	5.00E-25	390	42
TC22	hypothetical protein DDBDRAFT_0188625 [Dictyostelium discoideum AX4] (gi|66804457)-Dictyostelium discoideum AX4	516	YES	9.00E-34	1143	33
TC1416	unnamed protein product [Candida glabrata] (gi|50290703)-Candida glabrata CBS 138	480	YES	1.00E-25	1553	31
TC20	unnamed protein product [Candida glabrata] (gi|50290703)-Candida glabrata CBS 138	527	YES	5.00E-36	1553	35
TC1283	Os03g0309300 [Oryza sativa (japonica cultivar-group)] (gi|115452621)-Oryza sativa (japonica cultivar-group)	452	YES	4.00E-92	590	48
TC110	dragline silk protein (gi|107124034)-synthetic construct	519	YES	1.00E-87	528	40
TC111	dragline silk protein (gi|107124034)-synthetic construct	487	YES	7.00E-81	528	40
TC1569	keratin 9 (gi|148670627)-Mus musculus	224	Yes	1.00E-19	743	38
TC368	flagelliform silk protein (gi|7106228)-Nephila madagascariensis	221	Yes	1.00E-21	1884	41
TC183	PREDICTED: hypothetical protein (gi|110761334)-Apis mellifera	248	Yes	4.00E-23	608	36
TC182	glycine-rich protein [Arabidopsis thaliana] (gi|2961347)-Arabidopsis thaliana	242	Yes	4.00E-23	396	38
TC2299	flagelliform silk protein (gi|7106224)-Nephila clavipes	237	Yes	8.00E-26	2249	41
CF972311	flagelliform silk protein (gi|13561980)-Argiope trifasciata	239	Yes	4.00E-22	651	43
CF809432	EBNA-1 [Human herpesvirus 4 type 1] (gi|82503233)-BAC cloning vector pEBAC190G	256	Yes	4.00E-21	641	39
TC546	Os03g0309300 [Oryza sativa (japonica cultivar-group)] (gi|115452621)-Oryza sativa (japonica cultivar-group)	254	Yes	2.00E-23	590	39
TC1847	flagelliform silk protein (gi|13561980)-Argiope trifasciata	261	Yes	3.00E-28	651	40
CD792000	flagelliform silk protein (gi|7106228)-Nephila madagascariensis	267	Yes	7.00E-29	1884	40
CD784774	flagelliform silk protein (gi|7106229)-Nephila madagascariensis	166	Yes	1.00E-14	626	42
TC594	hypothetical protein DDBDRAFT_0188625 [Dictyostelium discoideum AX4] (gi|66804457)-Dictyostelium discoideum AX4	446	Yes	1.00E-11	1143	27
TC360	PREDICTED: hypothetical protein (gi|109504489)-Rattus norvegicus	445	Yes	2.00E-12	507	27

© University of Pretoria

RaGI ID	Best match in GenBank[a]	Seq len.[c]	Signal Peptide present?	E-value	Len. of best match[b]	% identity
TC796	PREDICTED: hypothetical protein (gi|109504489)-Rattus norvegicus	441	Yes	2.00E-08	507	28
TC345	PREDICTED: similar to Ras-related GTP-binding protein ragA (gi|110764304)-Apis mellifera	316	Yes	1.00E-150	306	86
TC1303	flagelliform silk protein-1 (gi|47606845)-Araneus ventricosus	308	Yes	9.00E-27	563	54
TC389	unknown protein (gi|42572805)-Arabidopsis thaliana	375	Yes	5.00E-44	579	40
TC152	hypothetical protein OsJ_010161 (gi|125586014)-Oryza sativa (japonica cultivar-group)	380	Yes	5.00E-47	590	38
CF809390	fibroin-4 gi|1589022|prf||2209442B fibroin:ISOTYPE=4 (gi|1263289)-	207	Yes	1.00E-13	410	32
TC1326	PREDICTED: similar to mucin 19 (gi|113423316)-Homo sapiens	300	Yes	1.00E-08	7328	29
TC609	PE-PGRS family protein [Mycobacterium tuberculosis H37Ra] (gi|148663375)-Mycobacterium tuberculosis H37Ra	310	Yes	1.00E-14	717	33
CF809392	hypothetical protein OsI_025258 (gi|125558490)-Oryza sativa (indica cultivar-group)	101	Yes	1.00E-10	356	45
TC23	PT repeat family protein [Trichomonas vaginalis G3] (gi|123478387)-Trichomonas vaginalis G3	250	No	2.00E-16	607	42
TC208	20/24 kDa immunodominant saliva protein (gi|28932710)-Rhipicephalus appendiculatus	194	Yes	1.00E-63	195	68
CD794666	Collagen triple helix repeat [Mycobacterium vanbaalenii PYR-1] (gi|120402275)-Mycobacterium vanbaalenii PYR-1	284	Yes	1.00E-54	796	49
TC1268	shematrin-4 (gi|93102305)-Pinctada fucata	227	Yes	2.00E-42	306	51
TC1281	PREDICTED: hypothetical protein (gi|91082979)-Tribolium castaneum	250	No	1.00E-58	592	49
TC1699	unknown (gi|45479213)-Rhipicephalus haemaphysaloides haemaphysaloides	315	Yes	2.00E-39	506	36
TC1274	putative glycine-rich protein [Arabidopsis thaliana] (gi|3892703)-Arabidopsis thaliana	328	No	2.00E-55	608	45
TC2272	Hypothetical protein CBG15737 (gi|39588898)-Caenorhabditis briggsae	166	Yes	2.00E-39	259	63
TC1487	hypothetical protein OsJ_010161 (gi|125586014)-Oryza sativa (japonica cultivar-group)	535	Yes	1.00E-103	590	48
CD795856	Os03g0309300 [Oryza sativa (japonica cultivar-group)] (gi|115452621)-Oryza sativa (japonica cultivar-group)	269	Yes	1.00E-43	590	49
TC438	Os07g0440100 [Oryza sativa (japonica cultivar-group)] (gi|115471843)-Oryza sativa (japonica cultivar-group)	288	No	2.00E-29	422	38
CD792574	Glycine-rich cell wall structural protein precursor gi|7670029|dbj|BAA94983.1| (gi|27735191)-Arabidopsis thaliana	209	Yes	2.00E-32	349	40
TC1989	At5g46730 [Arabidopsis thaliana] (gi|45935055)-Arabidopsis thaliana	145	Yes	2.00E-17	270	47

© University of Pretoria

Appendix 1. Table A1.2

RaGI ID	Best match in GenBank[a]	Seq len.[c]	Signal Peptide present?	E-value	Len. of best match[b]	% identity
CD789766	IP13040p (gi|66771909)-Drosophila melanogaster	268	Yes	3.00E-32	538	40
TC639	CG15597-PA [Drosophila melanogaster] (gi|24644552)-Drosophila melanogaster	186	No	1.00E-17	171	55
CD795293	hypothetical protein [Paramecium tetraurelia] (gi|145542500)-Paramecium tetraurelia	274	Yes	2.00E-24	288	43
Mucins						
TC2039	hypothetical protein LOC663263 [Tribolium castaneum] (gi|121583754)-Tribolium castaneum	247	Yes	9.00E-84	274	57
TC1531	7DB family (gi|114153212)-Argas monolakensis	422	Yes	0.6	167	30
TC1296	hypothetical protein, unknown function (gi|68128325)-Leishmania major	163	Yes	2.00E-12	841	36
TC1290	PREDICTED: hypothetical protein, partial (gi|115930826)-Strongylocentrotus purpuratus	161	Yes	1.00E-12	202	46
TC1292	PREDICTED: hypothetical protein, partial (gi|115930826)-Strongylocentrotus purpuratus	161	Yes	7.00E-13	202	51
CD789893	F56H9.6 [Caenorhabditis elegans] (gi|71996221)-Caenorhabditis elegans	166	Yes	2.00E-12	245	36
TC1295	hypothetical protein, unknown function (gi|68128325)-Leishmania major	164	Yes	8.00E-12	841	36
TC1290	PREDICTED: hypothetical protein, partial (gi|125804565)-Danio rerio	166	Yes	6.00E-12	1059	37
TC1296	MotA/TolQ/ExbB proton channel [Paracoccus denitrificans PD1222] (gi|119383479)-Paracoccus denitrificans PD1222	157	Yes	2.00E-10	431	34
TC1294	F56H9.6 [Caenorhabditis elegans] (gi|71996221)-Caenorhabditis elegans	164	Yes	1.00E-11	245	36
TC1291	glycoprotein gp2 (gi|17221104)-Equine herpesvirus 1	165	Yes	1.00E-11	337	35
TC490	visgun CG16707-PA, isoform A [Drosophila melanogaster] (gi|24661856)-Drosophila melanogaster	142	Yes	1.00E-25	182	54
CD797083	hypothetical protein EhV364 [Emiliania huxleyi virus 86] (gi|73852838)-Emiliania huxleyi virus 86	185	Yes	4.00E-32	2332	52
CD783057	hypothetical protein DDBDRAFT_0186922 [Dictyostelium discoideum AX4] (gi|66808255)-Dictyostelium discoideum AX4	151	Yes	3.00E-22	1473	68
CD785323	Hypothetical protein CBG11288 (gi|39590335)-Caenorhabditis briggsae	235	Yes	1.00E-12	502	32
CD790167	C30H6.11 [Caenorhabditis elegans] (gi|133900667)-Caenorhabditis elegans	216	Yes	1.00E-17	460	32
CD790740	PREDICTED: hypothetical protein (gi|94404358)-Mus musculus	234	Yes	4.00E-15	445	38
TC2129	PREDICTED: similar to TPRXL protein (gi|114585518)-Pan troglodytes	223	Yes	7.00E-07	215	25
Antigen 5 family						

© University of Pretoria

RaGI ID	Best match in GenBank[a]	Seq len.[c]	Signal Peptide present?	E-value	Len. of best match[b]	% identity
TC1264	ENSANGP00000010876 [Anopheles gambiae str. PEST] (gi|58378175)-Anopheles gambiae str. PEST	289	Yes	1.00E-23	146	41
TC919	SCP-related protein [Bombyx mori] (gi|148298863)-Bombyx mori	293	Yes	2.00E-24	371	38
TC115	F49E11.5 [Caenorhabditis elegans] (gi|17540532)-Caenorhabditis elegans	294	Yes	6.00E-19	212	34
CD796020	salivary gland protein (gi|50363178)-Ixodes scapularis	271	Yes	1.00E-52	320	55
Prokineticin domain family						
CD785689	Is6 (gi|29649848)-Ixodes scapularis	123	Yes	6.00E-06	115	40
TC2382	putative secreted salivary gland protein (gi|67083130)-Ixodes scapularis	132	Yes	3.00E-09	125	34
TC1822	Astakine precursor gi|60202501|gb|AAX14636.1| astakine (gi|74897764)-Penaeus monodon	127	Yes	4.00E-14	124	39
Proteins containing protease inhibitor domains						
TC1061	PREDICTED: similar to Bikunin hlg (gi|118099905)-Gallus gallus	77	No	0.005	570	32
CF809387	chymotrypsin inhibitor preproprotein (gi|37788275)-Vipera ammodytes	94	Yes	2.00E-09	93	36
CD792029	ZK287.4 [Caenorhabditis elegans] (gi|17566852)-Caenorhabditis elegans	142	Yes	0.19	1208	30
TC1823	putative salivary protein with Kunitz domains (gi|67083461)-Ixodes scapularis	148	No	7.00E-07	324	30
TC761	putative salivary protein with Kunitz domains (gi|67083461)-Ixodes scapularis	150	Yes	5.00E-12	324	34
CD790025	putative secreted protein (gi|76786687)-Boophilus microplus	181	No	2.00E-15	353	42
CD791005	hypothetical protein (gi|112359380)-Spironucleus barkhanus	253	Yes	0.003	858	25
TC134	Kunitz-like protease inhibitor precursor (gi|22901764)-Ancylostoma caninum	409	No	2.00E-81	759	39
TC527	Kunitz-like protease inhibitor precursor (gi|22901764)-Ancylostoma caninum	281	Yes	2.00E-63	759	42
TC620	PREDICTED: similar to Papilin CG33103-PB, isoform B isoform 1 (gi|110756487)-Apis mellifera	302	Yes	2.00E-17	2807	31
CD791646	Hypothetical protein CBG20566 (gi|39581092)-Caenorhabditis briggsae	273	Yes	4.00E-33	2157	39
TC1129	PREDICTED: hypothetical protein (gi|113425037)-Homo sapiens	352	Yes	5.00E-06	413	28
CD785622	conserved hypothetical protein (gi|108882055)-Aedes aegypti	51	Yes	14	537	38
CD796670	CG1637-PA, isoform A [Drosophila melanogaster] (gi|24641132)-Drosophila melanogaster	287	Yes	3.00E-91	453	57
TC443	transducin (beta)-like 2 [Xenopus laevis] (gi|147905556)-Xenopus laevis	236	Yes	3.00E-12	436	35
TC1740	PREDICTED: similar to CG16791-PA (gi|110763910)-Apis mellifera	240	Yes	4.00E-62	293	51
CD788731	predicted protein [Ostreococcus lucimarinus CCE9901] (gi|145345300)-Ostreococcus lucimarinus CCE9901	210	Yes	0.003	1451	34

© University of Pretoria

Appendix 1. Table A1.2

RaGI ID	Best match in GenBank[a]	Seq len.[c]	Signal Peptide present?	E-value	Len. of best match[b]	% identity
CD780808	hypothetical protein Dgeo_1645 [Deinococcus geothermalis DSM 11300] (gi|94985745)-Deinococcus geothermalis DSM 11300	222	Yes	0.005	280	28
TC127	serpin-1 precursor (gi|115334935)-Ixodes ricinus	399	Yes	5.00E-95	392	46
TC1726	serpin-8 precursor (gi|115334941)-Ixodes ricinus	403	Yes	0	402	79
TC1759	serine proteinase inhibitor serpin-2 (gi|17223664)-Rhipicephalus appendiculatus	377	No	0	380	90
TC815	hypothetical protein (gi|14140097)-Ixodes ricinus	361	No	7.00E-91	377	50
TC1592	serine proteinase inhibitor serpin-3 (gi|17223666)-Rhipicephalus appendiculatus	399	Yes	0	398	96
TC1528	serine proteinase inhibitor serpin-4 (gi|17223668)-Rhipicephalus appendiculatus	391	Yes	1.00E-143	485	71
CD792077	serine proteinase inhibitor serpin-4 (gi|17223668)-Rhipicephalus appendiculatus	271	Yes	4.00E-61	485	49
CD783353	serine proteinase inhibitor serpin-2 (gi|17223664)-Rhipicephalus appendiculatus	188	No	2.00E-70	380	80
TC2012	putative thyropin precursor (gi|41352539)-Ornithodoros moubata	270	Yes	5.00E-20	126	43
TC1206	conserved hypothetical protein (gi|108872476)-Aedes aegypti	215	Yes	4.00E-13	249	29
CD784209	PREDICTED: similar to Y69H2.3a isoform 1 (gi|91080607)-Tribolium castaneum	187	No	2.00E-12	199	32
TC2089	Y69H2.3c [Caenorhabditis elegans] (gi|32566734)-Caenorhabditis elegans	164	Yes	1.00E-16	731	37
CD788067	von Willebrand factor (gi|33285889)-Ixodes ricinus	93	Yes	5.00E-08	136	44
CF809396	Hypothetical protein CBG19173 (gi|39594507)-Caenorhabditis briggsae	144	Yes	3.00E-12	141	35
CD783893	unknown (gi|19343413)-Ectocarpus siliculosus virus	187	Yes	4.00E-05	590	25
CD784641	Translocated actin-recruiting phosphoprotein (Tarp protein) (gi|62901082)-Chlamydia trachomatis	196	Yes	9.00E-05	1005	28
CD796855	hypothetical protein, conserved (gi|68127008)-Leishmania major	184	Yes	6.00E-08	831	32
Basic Tail family						
TC766	BTSP-6 (gi|51011462)-Ixodes pacificus	153	Yes	3.00E-08	125	32
CF809389	putative salivary secreted protein with basic tail (gi|67083723)-Ixodes scapularis	115	Yes	1.00E-04	121	29
TC929	hypothetical protein LOC566596 [Danio rerio] (gi|113677023)-Danio rerio	213	Yes	2.00E-32	1001	37
CD782695	procyclic form surface glycoprotein [Trypanosoma brucei TREU927] (gi|71748552)-Trypanosoma brucei	138	Yes	1.00E-07	425	34
CF809408	hypothetical protein PY01301 [Plasmodium yoelii yoelii str. 17XNL] (gi|82915266)-Plasmodium yoelii yoelii	106	Yes	0.003	465	36
TC2380	PREDICTED: hypothetical protein [Strongylocentrotus purpuratus] (gi|115753209)-Strongylocentrotus purpuratus	235	Yes	1.00E-09	147	53

© University of Pretoria

RaGI ID	Best match in GenBank[a]	Seq len.[c]	Signal Peptide present?	E-value	Len. of best match[b]	% identity
TC1953	H39E23.3 [Caenorhabditis elegans] (gi|71997558)-Caenorhabditis elegans	221	Yes	1.00E-07	746	27
TC161	hypothetical protein, unknown function (gi|68128325)-Leishmania major	422	Yes	0.004	841	35
Lipocalin superfamily						
TC994	Female-specific histamine-binding protein 2 precursor (FS-HBP2) (gi|8470378)-Rhipicephalus appendiculatus	177	No	1.00E-96	190	93
CD789574	Male-specific histamine-binding salivary protein precursor (MS-HBP) (gi|8470381)-Rhipicephalus appendiculatus	199	Yes	2.00E-20	200	33
TC1138	hypothetical protein lpl1679 [Legionella pneumophila str. Lens] (gi|54294603)-Legionella pneumophila str. Lens	192	Yes	0.73	276	21
CD784141	hypothetical protein TP01_0758 [Theileria parva strain Muguga] (gi|71033275)-Theileria parva	144	Yes	2.1	696	33
CD791328	hypothetical protein (gi|68051217)-Haemaphysalis longicornis	196	Yes	7.00E-07	178	26
CD788077	hypothetical protein (gi|68051217)-Haemaphysalis longicornis	196	Yes	3.00E-04	178	28
TC1235	hypothetical protein (gi|68051217)-Haemaphysalis longicornis	199	Yes	2.00E-05	178	24
TC346	NADH dehydrogenase (quinone) [Pelobacter propionicus DSM 2379] (gi|118581194)-Pelobacter propionicus DSM 2379	193	Yes	4.8	633	23
TC2494	bifunctional cbiH protein and precorrin-3B C17-methyltransferase (gi|123969261)-Prochlorococcus marinus str. AS9601	206	Yes	1.9	600	22
CD795421	hypothetical protein DEHA0B05830g [Debaryomyces hansenii CBS767] (gi|50413185)-Debaryomyces hansenii CBS767	204	Yes	0.65	415	25
TC1154	hypothetical protein [Paramecium tetraurelia] (gi|145498383)-Paramecium tetraurelia	192	Yes	0.43	538	32
TC1624	Blo t profilin allergen (gi|33667952)-Blomia tropicalis	131	Yes	6.00E-50	130	71
TC2294	hypothetical protein DDBDRAFT_0218235 [Dictyostelium discoideum AX4] (gi|66814406)-Dictyostelium discoideum AX4	248	Yes	0.005	1071	25
TC1143	putative secreted histamine binding protein of 19.7 kDa (gi|51011480)-Ixodes pacificus	272	Yes	2.00E-08	196	29
TC1593	putative serotonin and histamine binding protein (gi|82791912)-Rhipicephalus haemaphysaloides haemaphysaloides	217	No	1.00E-46	219	48
TC410	putative serotonin and histamine binding protein (gi|82791912)-Rhipicephalus haemaphysaloides haemaphysaloides	217	Yes	5.00E-66	219	59
CD782167	putative serotonin and histamine binding protein (gi|82791912)-Rhipicephalus haemaphysaloides haemaphysaloides	214	No	9.00E-65	219	57

© University of Pretoria

Appendix 1. Table A1.2

RaGI ID	Best match in GenBank[a]	Seq len.[c]	Signal Peptide present?	E-value	Len. of best match[b]	% identity
TC780	P29 (gi|55824387)-Hyalomma asiaticum asiaticum	271	Yes	9.00E-09	254	22
Ixodidae 8.9 kDa family						
TC1539	putative secreted salivary protein (gi|67083146)-Ixodes scapularis	111	Yes	2.00E-04	105	34
TC278	putative 8.9 kDa secreted protein (gi|22164312)-Ixodes scapularis	111	Yes	2.00E-04	104	30
CD789474	putative 8.9 kDa secreted protein (gi|22164312)-Ixodes scapularis	111	Yes	2.00E-04	104	30
TC1250	putative secreted salivary protein (gi|67083146)-Ixodes scapularis	111	Yes	3.00E-04	105	34
CD784666	putative secreted salivary protein (gi|67083553)-Ixodes scapularis	104	Yes	0.71	101	33
TC1023	unnamed protein product (gi|47224127)-Tetraodon nigroviridis	174	Yes	2.00E-04	1215	30
TC898	unnamed protein product (gi|47224127)-Tetraodon nigroviridis	187	Yes	0.001	1215	30
Enzymes						
TC388	salivary gland metalloprotease (gi|71726988)-Boophilus microplus	521	Yes	1.00E-133	492	48
TC344	salivary gland metalloprotease (gi|71726988)-Boophilus microplus	522	Yes	1.00E-130	492	47
TC1474	metalloprotease (gi|84570466)-Haemaphysalis longicornis	478	Yes	1.00E-120	482	45
TC1404	metalloprotease (gi|122080312)-Haemaphysalis longicornis	493	Yes	1.00E-125	397	58
TC1409	salivary gland metalloprotease (gi|71726992)-Boophilus microplus	546	Yes	1.00E-127	506	46
TC1474	truncated secreted metalloprotease (gi|22164294)-Ixodes scapularis	246	No	4.00E-48	367	40
CD786875	salivary gland metalloprotease (gi|71726992)-Boophilus microplus	188	No	3.00E-72	506	64
CD782898	metalloprotease (gi|122080312)-Haemaphysalis longicornis	217	No	2.00E-55	397	68
TC44	salivary gland metalloprotease (gi|71726986)-Boophilus microplus	492	Yes	0	493	87
TC43	salivary gland metalloprotease (gi|71726986)-Boophilus microplus	492	Yes	0	493	86
TC45	salivary gland metalloprotease (gi|71726986)-Boophilus microplus	495	Yes	0	493	84
TC1311	salivary gland metalloprotease (gi|71726990)-Boophilus microplus	335	No	1.00E-106	559	57
TC1310	salivary gland metalloprotease (gi|71726990)-Boophilus microplus	307	No	1.00E-104	559	58
TC1309	salivary gland metalloprotease (gi|71726990)-Boophilus microplus	294	No	1.00E-102	559	59
TC34	metalloprotease (gi|121308309)-Haemaphysalis longicornis	308	No	4.00E-67	550	43
TC57	salivary gland metalloprotease (gi|71726984)-Boophilus microplus	477	Yes	0	478	84
TC1548	metalloproteinase (gi|89277230)-Rhipicephalus haemaphysaloides	475	Yes	0	468	82
TC1702	metalloproteinase (gi|89277230)-Rhipicephalus haemaphysaloides	473	Yes	1.00E-179	468	63

© University of Pretoria

RaGI ID	Best match in GenBank[a]	Seq len.[c]	Signal Peptide present?	E-value	Len. of best match[b]	% identity
TC1714	salivary gland metalloprotease (gi|71726984)-Boophilus microplus	476	Yes	1.00E-122	478	48
TC1857	salivary gland metalloprotease (gi|71726984)-Boophilus microplus	489	Yes	1.00E-125	478	47
CD788610	metalloproteinase (gi|89277230)-Rhipicephalus haemaphysaloides	281	Yes	7.00E-87	468	59
CD793291	metalloprotease (gi|122080312)-Haemaphysalis longicornis	200	No	3.00E-32	397	41
TC495	metalloprotease (gi|122080312)-Haemaphysalis longicornis	369	Yes	3.00E-40	397	37
CD793290	metalloprotease (gi|114153162)-Argas monolakensis	268	No	6.00E-37	273	34
CD796220	metalloprotease (gi|84570462)-Haemaphysalis longicornis	315	Yes	4.00E-35	506	29
CD778393	metalloprotease (gi|84570462)-Haemaphysalis longicornis	259	No	2.00E-29	506	32
TC1163	salivary gland metalloprotease (gi|71726986)-Boophilus microplus	485	Yes	9.00E-52	493	31
CD785346	metalloproteinase (gi|89277230)-Rhipicephalus haemaphysaloides	195	Yes	1.00E-50	468	55
CD784037	metalloprotease (gi|122080312)-Haemaphysalis longicornis	193	No	4.00E-21	397	33
CD787220	metalloprotease (gi|114153162)-Argas monolakensis	147	No	9.00E-17	273	36
TC2221	metalloprotease (gi|84570468)-Haemaphysalis longicornis	223	Yes	1.00E-06	274	31
TC133	metalloprotease (gi|84570468)-Haemaphysalis longicornis	381	Yes	2.00E-10	274	24
TC1799	salivary gland metalloprotease (gi|71726990)-Boophilus microplus	476	Yes	2.00E-36	559	26
TC1695	metalloproteinase (gi|89277230)-Rhipicephalus haemaphysaloides	307	No	7.00E-18	468	25
CD789200	metalloprotease (gi|84570466)-Haemaphysalis longicornis	300	No	2.00E-21	482	24
TC248	factor D-like protein (gi|27466898)-Dermacentor andersoni	373	Yes	0	375	95
TC1656	fed tick salivary protein 10 (gi|55736035)-Ixodes scapularis	392	Yes	1.00E-122	394	55
CD783756	CG3355-PA, isoform A [Drosophila melanogaster] (gi|24581698)-Drosophila melanogaster	112	No	1.00E-14	314	51
TC1075	PREDICTED: similar to serine carboxypeptidase vitellogenic-like (gi|73976541)-Canis familiaris	193	Yes	6.00E-24	479	36
CD796802	Probable serine carboxypeptidase CPVL precursor gi|55725296|emb|CAH89513.1| (gi|68565026)-Pongo pygmaeus	227	Yes	1.00E-32	476	43
CD787447	Probable serine carboxypeptidase CPVL precursor gi|55725296|emb|CAH89513.1| (gi|68565026)-Pongo pygmaeus	200	Yes	7.00E-23	476	40
CD783231	PREDICTED: similar to Carboxypeptidase, vitellogenic-like (gi|126341796)-Monodelphis domestica	162	Yes	2.00E-16	752	48
CD793973	putative secreted carboxypeptidase (gi|22164290)-Ixodes scapularis	215	No	3.00E-39	350	41

© University of Pretoria

Appendix 1. Table A1.2

RaGI ID	Best match in GenBank[a]	Seq len.[c]	Signal Peptide present?	E-value	Len. of best match[b]	% identity
CD784466	serine carboxypeptidase 1 (gi|71841605)-Triatoma infestans	170	Yes	4.00E-22	474	46
CD785353	putative secreted carboxypeptidase (gi|22164290)-Ixodes scapularis	206	Yes	8.00E-26	350	40
CD778664	Prcp protein (gi|20072291)-Mus musculus	247	Yes	2.00E-73	451	56
TC1761	esterase [Boophilus microplus] (gi|6003567)-Rhipicephalus microplus	282	No	1.00E-154	544	92
TC1015	esterase [Boophilus microplus] (gi|6003567)-Rhipicephalus microplus	253	No	2.00E-64	544	46
TC2499	esterase [Boophilus microplus] (gi|6003567)-Rhipicephalus microplus	262	Yes	9.00E-77	544	59
TC2347	esterase [Boophilus microplus] (gi|6003567)-Rhipicephalus microplus	272	Yes	1.00E-63	544	53
CD796734	esterase [Boophilus microplus] (gi|6003567)-Rhipicephalus microplus	277	Yes	1.00E-97	544	63
CD789591	acetylcholinesterase 3 AChE3 (gi|32966205)-Boophilus microplus	285	No	1.00E-55	620	41
CD786688	chitinase 7 [Tribolium castaneum] (gi|110431374)-Tribolium castaneum	246	No	8.00E-35	980	33
TC1483	PREDICTED: similar to MGC84097 protein (gi|50747900)-Gallus gallus	392	Yes	1.00E-103	393	49
CD790646	chitinase (gi|23956481)-Araneus ventricosus	200	Yes	4.00E-49	431	50
TC2063	putative salivary secreted peptide (gi|51011418)-Ixodes pacificus	197	Yes	6.00E-76	196	65
TC76	phospholipase A2 (gi|114153140)-Argas monolakensis	406	Yes	4.00E-28	221	42
CD788621	Sphingomyelin phosphodiesterase D precursor (Sphingomyelinase D) (SMD) (SMase D) (gi|121962650)-Ixodes scapularis	242	Yes	3.00E-50	364	51
CD790143	Sphingomyelin phosphodiesterase D precursor (Sphingomyelinase D) (SMD) (SMase D) (gi|121962650)-Ixodes scapularis	187	No	1.00E-43	364	49
CD781456	sphingomyelin phosphodiesterase, acid-like 3B [Rattus norvegicus] (gi|71043890)-Rattus norvegicus	253	Yes	2.00E-31	456	34
CD796848	PREDICTED: similar to neutral sphingomyelinase 3, partial [Strongylocentrotus (gi|115717537)-Strongylocentrotus purpuratus	200	No	2.00E-25	874	35
CD795896	Protein 5NUC precursor [Includes: UDP-sugar hydrolase (UDP-sugar diphosphatase) (gi|12229680)-Lutzomyia longipalpis	252	No	4.00E-48	572	40
CD784830	PREDICTED: similar to CG30104-PA, isoform A (gi|66523706)-Apis mellifera	223	Yes	9.00E-42	593	40
CD782471	79 kDa salivary apyrase precursor (gi|34481604)-Triatoma infestans	207	Yes	2.00E-33	557	43
CD791733	PREDICTED: similar to CG11883-PA, isoform A (gi|91086067)-Tribolium castaneum	235	No	2.00E-84	645	64
CD792896	PREDICTED: similar to 5-nucleotidase, ecto (CD73), partial [Strongylocentrotus (gi|115643671)-Strongylocentrotus purpuratus	71	Yes	2.00E-09	117	49

© University of Pretoria

RaGI ID	Best match in GenBank[a]	Seq len.[c]	Signal Peptide present?	E-value	Len. of best match[b]	% identity
CD790008	ectonucleotide pyrophosphatase/phosphodiesterase 6 [Xenopus laevis] (gi|148228249)-Xenopus laevis	287	Yes	1.00E-55	441	43
TC1668	PREDICTED: similar to CG4123-PA, isoform A (gi|91087961)-Tribolium castaneum	474	Yes	1.00E-43	731	30
TC712	hypothetical protein LOC641570 [Danio rerio] (gi|83025082)-Danio rerio	175	Yes	2.00E-36	341	52
TC400	angiotensin-converting enzyme-like protein (gi|1468981)-	273	Yes	1.00E-144	660	92
CD789984	angiotensin converting enzyme (gi|40365371)-Locusta migratoria	225	Yes	8.00E-28	625	39
TC2487	PREDICTED: alkaline phosphatase, liver/bone/kidney (gi|118101263)-Gallus gallus	216	Yes	1.00E-43	347	46
CD782583	PREDICTED: alkaline phosphatase, liver/bone/kidney (gi|118101263)-Gallus gallus	183	Yes	2.00E-40	347	50
CD781559	alkaline phosphatase, liver/bone/kidney [Bos taurus] (gi|110347574)-Bos taurus	202	Yes	1.00E-42	524	50
CD791376	hypothetical protein CC1G_10842 (gi|116505174)-Coprinopsis cinerea okayama7#130	245	Yes	1.00E-09	1699	25
CD785321	Ribonuclease Oy (RNase Oy) (gi|47117147)-	157	Yes	5.00E-26	213	41
TC503	Deoxyribonuclease II beta [Homo sapiens] (gi|133778285)-Homo sapiens	217	Yes	1.00E-22	357	39
TC558	unnamed protein product (gi|90078238)-Macaca fascicularis	263	Yes	1.00E-24	361	32
TC138	Plancitoxin-1 precursor (Plancitoxin I) (Plan-I) [Contains: Plancitoxin-1 (gi|74797443)-Acanthaster planci	371	Yes	6.00E-37	358	29
CD791783	deoxyribonuclease II beta [Xenopus tropicalis] (gi|58332460)-Xenopus tropicalis	174	Yes	2.00E-19	340	38
Immunity related products						
CD789133	latrophilin-like protein AD variant (gi|55602993)-Musca domestica	183	Yes	5.00E-33	1787	46
TC1145	PREDICTED: similar to Epididymal secretory protein E1 precursor (Niemann Pick (gi|91081217)-Tribolium castaneum	178	Yes	2.00E-05	144	26
CD782765	alpha-2-macroglobulin (gi|22164280)-Ixodes scapularis	196	No	2.00E-78	390	70
TC1764	complement component 3-like protein (gi|33325642)-Carcinoscorpius rotundicauda	299	No	8.00E-38	1737	31
CD795445	complement component 3-like protein (gi|33325642)-Carcinoscorpius rotundicauda	304	No	4.00E-57	1737	40
TC1081	PREDICTED: similar to CG6038-PA (gi|91077808)-Tribolium castaneum	159	Yes	1.00E-35	155	47
CD782147	unnamed protein product (gi|47210688)-Tetraodon nigroviridis	75	Yes	0.25	1058	34
TC1977	microplusin preprotein-like (gi|67083118)-Ixodes scapularis	136	Yes	7.00E-06	125	31

© University of Pretoria

Appendix 1. Table A1.2

RaGI ID	Best match in GenBank[a]	Seq len.[c]	Signal Peptide present?	E-value	Len. of best match[b]	% identity
CD787867	conserved hypothetical protein [Aspergillus fumigatus Af293] (gi|70991258)-Aspergillus fumigatus Af293	54	Yes	52	339	42
CD794403	Y39B6A.1 [Caenorhabditis elegans] (gi|25151613)-Caenorhabditis elegans	159	Yes	2.00E-08	735	34
TC1768	Y39B6A.1 [Caenorhabditis elegans] (gi|25151613)-Caenorhabditis elegans	173	Yes	2.00E-07	735	32
TC69	S-antigen protein precursor gi|160065|gb|AAA29472.1| S-antigen precursor (gi|134206)-	252	Yes	3.00E-17	375	38
TC67	S-antigen protein precursor gi|160065|gb|AAA29472.1| S-antigen precursor (gi|134206)-	436	Yes	9.00E-37	375	37
Conserved Secreted proteins						
TC1739	PREDICTED: similar to CG11958-PB, isoform B (gi|91080995)-Tribolium castaneum	317	Yes	1.00E-106	585	64
TC2490	secreted protein (gi|67084071)-Ixodes scapularis	323	Yes	1.00E-142	324	80
TC58	unknown (gi|67083415)-Ixodes scapularis	144	Yes	8.00E-42	145	62
TC58	unknown (gi|67083415)-Ixodes scapularis	135	Yes	9.00E-39	145	63
TC136	salivary selenoprotein M precursor (gi|67083573)-Ixodes scapularis	127	Yes	6.00E-37	129	68
TC596	salivary selenoprotein precursor (gi|67083487)-Ixodes scapularis	152	No	3.00E-55	150	73
CD789378	putative salivary protein (gi|67083339)-Ixodes scapularis	231	Yes	1.00E-114	229	84
TC2016	transmembrane prostate androgen-induced protein [Gallus gallus] (gi|71895609)-Gallus gallus	197	Yes	1.00E-06	286	31
TC1988	PREDICTED: hypothetical protein [Strongylocentrotus purpuratus] (gi|115677355)-Strongylocentrotus purpuratus	122	Yes	4.00E-11	201	34
CD794512	PREDICTED: similar to nel-like 1 precursor (gi|91081523)-Tribolium castaneum	160	Yes	1.00E-12	874	39
CD782190	Niemann-Pick type C1 protein (gi|6934272)-Cricetulus griseus	192	Yes	1.00E-21	1277	34
CD779673	conserved hypothetical protein (gi|108882039)-Aedes aegypti	103	Yes	6.00E-22	148	46
CD783803	hypothetical protein LOC450038 [Danio rerio] (gi|54400618)-Danio rerio	252	Yes	4.00E-34	298	38
TC1127	PREDICTED: similar to CG31637-PA (gi|91091688)-Tribolium castaneum	383	Yes	3.00E-29	381	29
TC1840	conserved hypothetical protein (gi|108879860)-Aedes aegypti	263	Yes	9.00E-40	288	59
CD785128	LD22337p (gi|21744261)-Drosophila melanogaster	173	Yes	2.00E-25	524	39
CD796128	PREDICTED: similar to CG3655-PA (gi|91084727)-Tribolium castaneum	210	Yes	8.00E-18	541	45
TC1967	PREDICTED: hypothetical protein (gi|118094519)-Gallus gallus	250	Yes	3.00E-58	265	54

[a] BLASTX sequence search used
[b] len=length of amino acids

© University of Pretoria

Table A1.3 Properties of Tentative Consensus (TC) sequences in RaGI that lack significant matches with sequences in public databases.

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC1001	59.5461	749	2	154
TC1002	57.0166	905	2	259
TC1009	37.587	862	2	59
TC1010	42.2658	918	2	103
TC1012	68.5783	837	2	209
TC1016	47.6082	878	2	98
TC102	57.3913	920	14	223
TC1023	46.9441	769	2	174
TC1024	50.4425	791	2	231
TC1025	61.7409	988	2	170
TC1026	64.5374	908	2	209
TC1028	44.8925	744	2	88
TC1029	46.347	876	2	130
TC1033	46.723	946	2	105
TC1038	43.3581	941	2	86
TC1040	46.8487	952	2	165
TC1041	37.8652	890	2	71
TC1047	54.1489	940	2	295
TC1048	41.2141	939	2	65
TC1049	48.3573	974	2	221
TC1050	37.6096	912	2	67
TC1053	51.2221	941	2	112
TC1055	44.6985	962	2	91
TC1056	44.8718	936	2	125
TC1058	37.5916	955	2	85
TC1060	40.5006	879	2	67
TC1061	42.8101	911	2	113
TC1063	44.5853	868	2	76
TC1068	41.5865	832	2	103
TC1071	58.4009	863	2	259
TC1072	41.5778	938	2	108
TC1077	51.9565	920	2	140
TC1078	44.6352	932	2	94
TC1079	43.595	968	2	72
TC1080	53.7634	930	2	306
TC1087	41.3943	918	2	131
TC1088	48.1258	827	2	274
TC1089	45.1299	924	2	147
TC109	39.4231	2392	13	92
TC1090	56.4972	885	2	207
TC1091	42.1687	996	2	132

© University of Pretoria

Appendix 1. Table A1.3

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC1093	41.639	903	2	64
TC1097	48.4848	891	2	84
TC1098	42.7156	1009	2	106
TC1099	36.3273	1002	2	98
TC11	43.5132	3985	70	228
TC1101	45.6747	867	2	73
TC1104	56.1659	892	2	236
TC1105	49.8869	884	2	262
TC1108	49.6689	906	2	183
TC1109	42.0759	896	2	91
TC1110	34.9462	930	2	64
TC1113	60.2763	1158	2	223
TC1116	42.1053	893	2	79
TC1121	46.3527	1083	2	114
TC1123	47.4093	772	2	83
TC1124	57.265	1287	2	313
TC1125	38.0403	694	2	98
TC1126	39.403	670	2	63
TC1129	50.3766	1062	2	351
TC1130	42.7225	955	2	180
TC1132	57.6525	869	2	243
TC1138	41.2446	691	2	194
TC1140	53.7563	1198	2	214
TC1142	55.2511	657	2	138
TC1143	54.6148	1311	2	354
TC1150	44.5455	330	2	87
TC1151	38.4569	1646	2	92
TC1152	42.0261	997	2	128
TC1154	42.4337	641	2	192
TC1155	46.5294	1311	2	134
TC1156	38.1385	881	2	71
TC1161	39.4366	1349	2	72
TC1162	46.2783	1545	2	131
TC1164	36.9186	344	2	104
TC1165	35.4212	463	2	61
TC1167	57.5676	1480	2	345
TC1169	40.9677	310	2	61
TC1172	41.1817	1134	2	189
TC1173	46.4716	581	2	177
TC1174	42.711	782	2	124
TC1175	39.558	905	2	65
TC1178	49.8833	1285	2	300
TC1179	62.1936	775	2	257
TC1180	53.641	975	2	174
TC1182	45.0299	835	2	90

© University of Pretoria

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC1185	46.2357	1315	2	156
TC1186	42.1708	562	2	81
TC1189	43.5973	1023	2	149
TC119	47.4795	1706	12	176
TC1190	44.8575	807	2	88
TC1191	39.9302	859	2	62
TC1193	43.7363	910	2	73
TC1195	44.1739	1150	2	117
TC1200	49.0683	805	2	128
TC1202	51.8085	940	2	120
TC1203	41.4843	1051	2	135
TC1204	45.2719	846	2	91
TC1209	62.5536	932	2	270
TC1214	60.1369	1169	2	298
TC1216	39.2226	849	2	128
TC1217	41.4493	1035	2	108
TC1221	63.7255	918	2	209
TC1223	46.4941	927	2	86
TC1226	53.1915	846	2	132
TC123	44.2029	1794	11	198
TC1230	37.2922	842	2	80
TC1231	39.5474	928	2	104
TC1232	52.5093	1076	2	196
TC1233	43.2933	917	2	104
TC1235	40.4223	663	2	199
TC1237	44.0095	843	2	183
TC1238	46.5091	931	2	101
TC124	58.7286	991	11	171
TC1240	45.0382	1179	2	118
TC1242	40.5405	666	2	180
TC1247	42.7989	736	2	114
TC1252	42.2378	715	2	196
TC1256	44.9529	743	2	215
TC1258	40.8832	702	2	60
TC1259	40.2386	922	2	111
TC1260	47.1669	1306	2	179
TC1263	37.4491	737	2	76
TC1264	52.0408	1274	2	190
TC1266	44.469	904	2	106
TC1267	59.4961	516	2	104
TC1272	58.6611	1449	101	456
TC1274	56.8889	1800	83	568
TC1286	40.4812	1912	38	90
TC1287	38.746	622	3	82

© University of Pretoria

Appendix 1. Table A1.3

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC1288	39.0173	1038	9	85
TC1289	52.3892	2323	13	216
TC1290	59.3854	1204	12	227
TC1291	56.5848	1587	9	285
TC1292	59.778	1171	4	217
TC1293	40.625	2016	7	104
TC1294	57.784	1426	9	213
TC1295	53.8049	933	2	213
TC1296	59.2806	1390	7	213
TC1297	61.032	1124	2	209
TC1303	58.8159	1537	30	463
TC1308	54.3253	289	2	95
TC1313	49.0452	995	14	115
TC1314	49.1282	975	6	107
TC1317	44.7837	786	4	76
TC1323	45.5041	734	3	102
TC1324	49.2341	3656	33	1132
TC1325	48.0652	982	9	261
TC1326	50.592	929	3	299
TC1328	42.0019	1069	12	125
TC133	46.8389	1471	10	381
TC1330	40.5827	961	9	203
TC1331	49.3984	1579	35	422
TC1332	49.2769	968	5	264
TC1341	58.3741	1021	29	148
TC1344	43.6957	920	4	89
TC1347	43.0792	1351	6	164
TC1348	51.5123	1058	6	141
TC1349	51.6595	1868	13	229
TC1350	46.5608	945	3	83
TC1351	62.2568	1028	19	268
TC1352	61.461	794	2	246
TC1353	62.069	899	2	250
TC1354	50.6231	1685	23	166
TC1357	42.6386	1046	20	91
TC1358	42.3433	973	2	73
TC1378	42.6605	654	6	217
TC1380	49.0476	1890	17	236
TC1382	50.6809	1028	6	122
TC1385	47.9506	1537	16	479
TC1386	45.3381	1405	7	132
TC1397	44.7458	885	3	110
TC140	50.3715	673	2	89
TC1401	53.75	1360	14	268
TC1419	61.1386	808	2	241

© University of Pretoria

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC1425	42.651	1871	7	72
TC1426	45.1874	987	3	84
TC1432	41.8301	918	10	117
TC1434	42.6975	2595	10	103
TC1441	41.1591	1018	10	78
TC1442	41.9512	1845	10	168
TC1443	44.1939	887	10	112
TC1447	47.663	1626	9	206
TC145	42.2717	1708	8	115
TC1456	40.2762	869	3	66
TC1457	37.6707	1391	5	87
TC146	41.9244	873	2	78
TC1461	39.8374	861	3	83
TC1462	41.9689	1158	6	97
TC1467	44.9795	976	8	92
TC1471	41.5863	933	8	119
TC1472	43.8838	1308	8	87
TC1478	42.1161	2429	8	143
TC1481	42.4337	1923	8	270
TC1489	48.8682	2253	7	175
TC1490	44.9131	2703	6	101
TC1493	45.8772	1140	7	112
TC1496	43.4808	1695	6	87
TC1499	43.1698	1325	7	106
TC1504	44.7985	1067	3	89
TC1511	56.1392	1580	7	232
TC1515	44.1475	1085	6	73
TC1516	56.4009	1367	7	385
TC1523	49.9471	945	2	116
TC1525	62.2507	1404	6	303
TC1529	44.9561	912	6	142
TC1531	48.8087	1385	6	422
TC1532	40.3061	980	6	97
TC1533	38.6155	1849	6	106
TC1537	43.5227	1567	6	96
TC1538	43.4453	923	6	76
TC1541	39.2544	912	6	95
TC1546	38.7464	1053	6	67
TC155	39.14	907	9	91
TC1552	50.197	1777	6	109
TC1554	44.9198	1496	6	151
TC1562	51.3706	985	6	162
TC1563	44.1658	917	5	113
TC1564	43.7415	1470	5	143

© University of Pretoria

Appendix 1. Table A1.3

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC1569	56.0137	873	3	224
TC1570	55.7955	880	2	225
TC1573	48.5017	901	5	105
TC1575	46.6595	928	4	115
TC1576	38.0567	1235	5	68
TC1585	40.4086	881	5	107
TC1587	56.2651	830	5	154
TC1588	40.5615	1033	5	127
TC159	43.953	1447	9	185
TC1590	39.3554	2358	5	107
TC1594	40.7653	1359	5	94
TC1600	43.7079	890	5	145
TC1603	55.0388	903	5	268
TC161	54.5042	1321	9	386
TC1610	40.9533	1028	5	216
TC162	40.2633	1443	7	112
TC163	42.3709	852	2	105
TC1631	46.4024	1640	4	99
TC1632	48.6296	1277	4	405
TC1638	44.6965	1499	4	157
TC1641	46.317	896	4	116
TC1643	39.8681	1668	3	144
TC1644	54.902	867	4	193
TC1648	46.9449	671	4	122
TC1649	35.9189	838	3	99
TC165	40.339	1770	6	83
TC1650	39.1566	1494	4	145
TC1652	52.3688	781	4	112
TC1654	47.7684	829	4	188
TC1657	43.0861	2184	4	106
TC1659	40.0229	872	4	84
TC166	40.2899	1035	3	87
TC1660	43.1765	850	4	140
TC1662	45.0893	1344	4	98
TC1664	37.7801	937	4	94
TC1667	41.6851	902	4	142
TC167	45.3501	957	9	100
TC1670	46.5486	1637	4	117
TC1671	41.8745	1323	4	85
TC1684	41.6667	1044	4	89
TC1687	42.0827	701	4	116
TC1688	47.9361	751	2	129
TC1689	48.2419	711	2	126
TC1690	48.1395	1720	4	286
TC1697	60.2694	891	4	250

© University of Pretoria

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC17	42.0538	1227	10	110
TC1703	40.6593	1001	4	75
TC1704	41.7284	810	4	170
TC1705	54.2277	887	4	251
TC1708	49.3595	1327	3	145
TC171	42.8144	1002	6	86
TC1710	53.6913	894	4	168
TC1712	37.8558	1054	4	177
TC1719	40.265	981	4	93
TC1723	44.1116	968	4	107
TC1728	42.1336	928	4	94
TC1730	56.9832	1074	4	124
TC1738	40.1399	1430	4	89
TC1743	49.6538	1011	4	247
TC1745	62.7563	878	4	275
TC1748	44.9799	996	4	94
TC1752	38.9671	852	4	68
TC1753	52.6254	857	4	151
TC1754	41.6851	902	4	85
TC1765	54.5355	915	3	95
TC1766	48.7472	439	3	84
TC1771	47.9551	1516	3	125
TC1772	39.4413	895	3	104
TC1774	48.6604	1381	3	202
TC1778	39.1593	904	3	106
TC178	41.8182	935	5	92
TC1782	50.3257	1842	3	167
TC1789	39.624	1436	3	70
TC179	41.5571	989	3	92
TC1790	59.4017	936	3	183
TC1791	48.5868	743	3	97
TC1792	38.7097	1023	3	85
TC1796	43.9873	2212	3	127
TC1798	44.0294	1499	3	131
TC1802	41.4345	1687	3	92
TC1804	45.079	823	2	70
TC1805	44.3674	577	3	87
TC1806	39.8981	589	3	82
TC1807	35.8566	753	3	83
TC1808	54.6072	1031	3	213
TC1810	39.953	1702	3	123
TC1811	41.5753	1003	3	85
TC1816	44.6172	836	3	138
TC1819	42.4621	1121	3	87

© University of Pretoria

Appendix 1. Table A1.3

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC182	58.4633	898	3	242
TC1820	49.6257	935	3	100
TC1822	55.4801	1031	3	127
TC1823	43.5138	979	3	77
TC1824	42.2719	1074	3	95
TC1825	42.0922	1606	3	90
TC1826	56.2972	794	3	165
TC1827	45.1318	986	3	86
TC1828	42.0849	518	3	92
TC1831	55.4968	946	3	269
TC1834	48.7556	1567	3	187
TC1835	41.6935	933	3	94
TC184	58.2022	890	2	235
TC1842	43.9057	1485	2	117
TC1848	43.0131	916	3	163
TC1849	43.5982	906	3	70
TC1850	39.4883	899	3	62
TC1853	61.4679	872	3	221
TC1856	45.8937	1449	3	161
TC1863	38.4216	963	3	111
TC1864	33.7104	442	3	40
TC1869	44.2857	910	3	204
TC187	43.5851	2198	8	212
TC1874	38.4199	924	3	108
TC1878	48.0952	1260	3	344
TC188	50.1062	942	8	142
TC1883	42.7545	835	3	68
TC1884	43.8654	1516	3	113
TC1885	44.9573	585	3	125
TC1886	53.222	419	3	109
TC1887	44.5255	685	3	70
TC1888	37.7721	781	2	95
TC1892	46.0149	941	3	90
TC1893	45.9782	1007	3	156
TC1894	43.4088	751	3	75
TC1896	47.2755	1303	3	113
TC1903	40.8895	2091	3	109
TC1906	39.4028	1373	3	61
TC1909	40.5079	827	3	83
TC1912	73.1343	402	3	133
TC1914	42.9921	635	3	81
TC1918	40.4523	796	3	85
TC1919	44.5583	781	3	81
TC192	39.7804	1184	8	93
TC1920	50.2821	1418	3	131

© University of Pretoria

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC1923	46.8997	887	3	102
TC1924	41.5961	827	3	146
TC1929	36.0784	1020	3	58
TC1931	40.4545	660	3	80
TC1933	50.6122	980	3	105
TC1935	62.3681	853	3	147
TC1937	43.5402	983	3	117
TC1938	42.0323	866	3	138
TC1939	46.3675	468	2	148
TC1941	57.8411	1473	3	322
TC1942	50.495	808	2	91
TC1951	41.9126	847	3	150
TC1953	58.2979	940	3	217
TC1954	46.1538	1469	3	76
TC1956	50.6887	726	3	85
TC1957	37.1963	535	3	72
TC1959	59.125	800	3	199
TC1960	39.2461	902	3	62
TC1963	56.0897	936	3	259
TC1964	35.8403	1077	3	69
TC1969	41.9558	951	2	93
TC1971	46.6945	1437	2	117
TC1973	61.0245	449	2	148
TC1975	42.3529	680	2	80
TC1976	59.2299	883	2	228
TC1977	53.4413	494	2	136
TC1980	41.2668	1563	2	158
TC1982	40.3333	900	2	116
TC1985	59.798	495	2	132
TC1989	59.5661	507	2	145
TC1991	43.5811	1480	2	139
TC1992	57.529	518	2	123
TC1993	40.4375	1783	2	123
TC1995	58.1779	1394	2	380
TC1999	37.6157	864	2	77
TC200	53.3775	1510	7	183
TC2000	42.418	488	2	87
TC2001	40.3934	1525	2	90
TC2002	42.069	1305	2	93
TC2003	40.5861	1126	2	93
TC2004	46.9136	972	2	162
TC2007	37.741	726	2	82
TC2011	39.7328	1422	2	109
TC2014	55.1111	450	2	107

© University of Pretoria

Appendix 1. Table A1.3

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC2015	41.3744	1397	2	97
TC2017	46.5462	941	2	140
TC2018	49.5569	1354	2	305
TC2019	47.3214	784	2	134
TC2022	40.8609	1510	2	82
TC2023	39.7321	1120	2	130
TC2024	43.8335	1273	2	74
TC2025	43.2071	898	2	82
TC2027	45.2736	1407	2	113
TC2029	42.5754	862	2	118
TC2030	52.5151	497	2	122
TC2031	49.7836	924	2	131
TC2032	43.1034	1566	2	158
TC2033	41.067	806	2	84
TC2035	53.5112	712	2	232
TC2036	57.5163	612	2	144
TC2037	49.1176	680	2	91
TC2038	45.6332	916	2	146
TC204	37.5679	921	2	75
TC2040	40.7589	817	2	77
TC2043	59.2428	449	2	110
TC2045	39.2481	665	2	66
TC2046	45.5097	824	2	120
TC2048	57.3494	830	2	209
TC205	38.0355	957	2	75
TC2050	44.5476	431	2	73
TC2051	43.0905	1469	2	107
TC2052	47.076	342	2	84
TC2053	41.1834	1690	2	125
TC2056	55.2117	614	2	114
TC2058	44.5913	832	2	68
TC2059	37.9363	659	2	69
TC206	42.955	1022	3	126
TC2060	38.0757	977	2	65
TC2061	39.782	734	2	71
TC2065	41.9689	579	2	74
TC2068	46.9809	1358	2	135
TC2074	50.4082	1470	2	151
TC2077	52.1173	1535	2	272
TC2078	44.1176	1088	2	87
TC2079	39.4222	1073	2	119
TC2083	42.3168	423	2	68
TC2086	52.646	1455	2	128
TC2090	41.9922	1536	2	82
TC2091	40.7658	888	2	150

© University of Pretoria

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC2096	40.4056	1282	2	136
TC2097	57.6792	293	2	95
TC2098	45.8255	1593	2	108
TC2099	36.6667	1320	2	72
TC2100	36.3761	1457	2	79
TC2101	50.7511	932	2	130
TC2103	48.5484	620	2	95
TC2105	43.2854	834	2	69
TC2106	42.0395	1520	2	93
TC211	36.1771	926	7	111
TC2110	42.2572	381	2	104
TC2111	47.2296	758	2	128
TC2113	39.6936	1436	2	102
TC2114	45.9399	899	2	162
TC2117	39.4127	1294	2	86
TC2119	41.311	656	2	121
TC2121	53.019	1209	2	286
TC2123	51.6746	627	2	168
TC2124	43.8692	367	2	44
TC2126	40.3415	937	2	74
TC2128	46.0177	339	2	78
TC2129	65.4888	849	2	281
TC2130	40.5449	624	2	104
TC2133	43.9227	1448	2	91
TC2138	41.6058	1644	2	117
TC2140	44.5755	848	2	111
TC2141	37.5862	1160	2	107
TC2143	43.5556	900	2	115
TC2145	44.7975	1259	2	98
TC2148	48.0769	364	2	94
TC215	55.1791	1033	6	205
TC2150	45.7971	1380	2	342
TC2153	39.5876	485	2	110
TC2154	44.8819	1524	2	105
TC2155	35.0365	411	2	55
TC2156	41.0347	1411	2	124
TC2158	37.5962	1040	2	77
TC2160	55.7692	416	2	122
TC2162	42.4922	1605	2	117
TC2165	37.8132	878	2	82
TC2166	57.9853	407	2	125
TC2169	40.4387	1459	2	91
TC2170	42.9664	1699	2	110
TC2173	41.3454	1323	2	128

© University of Pretoria

Appendix 1. Table A1.3

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC2174	42.815	1016	2	76
TC2176	37.2593	883	2	72
TC2179	49.7722	878	2	221
TC2182	56.1724	1531	2	191
TC2183	41.292	1579	2	187
TC2184	60.4651	1634	2	301
TC2186	40.2089	383	2	90
TC2192	59.3081	607	2	118
TC2193	40.9685	1301	2	96
TC2194	57.2739	763	2	137
TC2195	33.4086	443	2	101
TC2196	41.5584	847	2	65
TC2197	43.7659	786	2	106
TC2198	45.6522	874	2	181
TC2199	41.6964	1403	2	78
TC220	40.707	877	4	96
TC2200	41.0681	543	2	75
TC2201	37.9845	1419	2	117
TC2202	44.1877	1385	2	121
TC2203	39.1965	921	2	83
TC2204	51.8152	303	2	78
TC2205	42.0345	1042	2	97
TC2206	53.7958	764	2	121
TC2207	39.5184	706	2	88
TC2209	37.477	1086	2	79
TC221	43.6563	1001	3	78
TC2211	37.5	736	2	82
TC2212	53.5789	950	2	242
TC2214	45.6221	1302	2	107
TC2215	45.4294	361	2	73
TC2216	57.7259	343	2	100
TC2218	57.56	1541	2	181
TC2220	62.9301	901	2	299
TC2221	46.5791	1447	2	223
TC2224	57.4257	808	2	251
TC2225	53.5242	908	2	172
TC2234	56.0227	880	2	195
TC2238	52.7363	804	2	194
TC224	59.9542	874	7	277
TC2240	44.9048	893	2	83
TC2242	55.8342	917	2	289
TC2244	48.6452	775	2	106
TC2246	46.5574	915	2	82
TC2251	43.9916	957	2	82
TC2253	58.3871	930	2	270

© University of Pretoria

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC2256	59.1176	1020	2	228
TC2258	40.5022	916	2	78
TC2259	45.2851	912	2	108
TC226	43.3781	1042	5	113
TC2260	39.1657	863	2	69
TC2261	38.5287	802	2	54
TC2263	52.9477	899	2	127
TC2266	47.4506	961	2	285
TC2268	44.098	898	2	70
TC227	38.676	287	2	57
TC2274	36.4883	729	2	77
TC2276	49.3075	722	2	129
TC2279	38.0282	923	2	88
TC228	40.6096	2428	7	90
TC2281	40.9171	1134	2	83
TC2284	40.4279	888	2	115
TC2285	35.7508	979	2	77
TC2288	45.0508	788	2	187
TC2290	56.5508	748	2	189
TC2292	50.6683	823	2	235
TC2297	39.9573	936	2	78
TC2298	63.4921	756	2	251
TC2299	56.9301	873	2	237
TC2303	52.3497	915	2	144
TC2304	40.959	1001	2	118
TC2308	38.4868	912	2	71
TC2309	36.1165	1030	2	68
TC2310	41.1125	827	2	69
TC2312	43.8127	897	2	135
TC2315	43.3962	954	2	162
TC2316	47.1344	1012	2	325
TC2319	56.9024	891	2	250
TC232	55.7403	1263	4	178
TC2320	45.2514	895	2	94
TC2321	45.2563	917	2	135
TC2322	35.6394	954	2	67
TC2325	49.5385	975	2	106
TC2328	48.8095	924	2	280
TC2329	40.02	1002	2	100
TC233	55.52	1250	3	164
TC2331	44.2607	1028	2	102
TC2337	41.8685	867	2	90
TC2338	45.8896	815	2	92
TC2339	42.0996	924	2	99

© University of Pretoria

Appendix 1. Table A1.3

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC2340	47.1554	914	2	91
TC2343	46.3597	934	2	110
TC2346	57.0312	896	2	231
TC2351	44.7072	888	2	86
TC2352	42.4211	950	2	171
TC2355	48.5588	902	2	203
TC2357	51.1879	926	2	124
TC2359	47.7547	913	2	90
TC2362	51.2118	949	2	141
TC2365	37.7232	896	2	65
TC2366	40.5433	994	2	104
TC2367	48.8837	851	2	251
TC2369	45.0624	881	2	89
TC2370	43.0296	911	2	117
TC2371	64.2202	872	2	184
TC2373	45.3465	1010	2	158
TC2374	51.0139	937	2	125
TC2377	43.5432	937	2	118
TC2380	39.0394	812	2	235
TC2382	44.6409	905	2	132
TC2387	42.7594	877	2	108
TC2388	41.5758	825	2	68
TC2392	39.8693	918	2	84
TC2393	35.7384	887	2	71
TC2395	51.5856	946	2	115
TC2396	41.3151	806	2	112
TC2397	39.1195	795	2	76
TC2399	45.3263	567	2	64
TC240	48.1679	1310	6	129
TC2401	38.1443	776	2	67
TC2402	40.4643	603	2	64
TC2406	52.3349	621	2	88
TC2411	40.4389	638	2	71
TC2412	38.5732	827	2	69
TC2417	61.3932	847	2	244
TC2418	41.2679	836	2	111
TC2419	57.7629	599	2	141
TC2424	55.202	817	2	159
TC2425	44.6626	815	2	71
TC2427	42.3077	312	2	93
TC2428	39.7794	1541	2	112
TC243	54.1422	1533	6	222
TC2430	46.7883	1261	2	113
TC2432	49.5789	950	2	154
TC2435	44.6878	913	2	69

© University of Pretoria

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC2438	40.1943	1647	2	97
TC2439	42.9087	832	2	92
TC2441	38.1579	304	2	77
TC2447	46.6055	545	2	101
TC2449	48.3834	866	2	219
TC2452	49.747	593	2	115
TC2455	37.594	798	2	128
TC2456	41.7629	953	2	92
TC2457	40.4088	636	2	66
TC2458	42.3319	952	2	98
TC2459	48.556	554	2	103
TC2460	61.3048	889	2	206
TC2461	53.8033	539	2	115
TC2463	47.1279	766	2	186
TC2464	47.2329	777	2	114
TC2466	55.4436	496	2	144
TC2467	40.4596	1436	2	125
TC2469	40.2023	791	2	79
TC2470	57.5832	811	2	218
TC2473	41.366	776	2	103
TC2475	42.4628	942	2	98
TC2479	37.7654	895	2	92
TC2482	46.5672	670	2	67
TC2483	45.3608	679	2	105
TC2485	46.9242	699	2	95
TC2488	39.5582	996	2	83
TC2491	46.6399	997	2	100
TC2493	47.8355	924	2	302
TC2494	43.2773	714	2	205
TC2496	37.5635	788	2	106
TC2497	45.6897	812	2	121
TC2502	50.7958	754	2	169
TC2503	57.1723	969	2	150
TC2504	46.0756	688	2	186
TC2510	45.7672	756	2	98
TC2511	41.3402	970	2	78
TC2512	50.815	1043	2	159
TC2515	41.4087	1292	2	61
TC2517	42.9854	891	2	76
TC2525	43.07	671	2	105
TC2526	48.4536	970	2	143
TC2528	46.6887	906	2	108
TC2531	58.0189	848	2	277
TC2533	38.5733	757	2	61

© University of Pretoria

Appendix 1. Table A1.3

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC2535	44.8878	802	2	90
TC2537	60.3306	1210	2	285
TC2539	43.4234	555	2	109
TC2540	41.033	697	2	101
TC2541	39.0428	794	2	86
TC266	43.1479	737	6	87
TC267	45.283	795	6	83
TC268	43.2602	957	3	72
TC269	42.6698	839	3	88
TC270	51.1435	1443	4	161
TC272	55.2555	1037	6	315
TC279	44.617	966	6	69
TC282	38.4106	906	5	161
TC286	51.3144	951	5	309
TC290	40.4079	1324	5	112
TC297	38.046	870	5	57
TC3	41.0383	3236	93	115
TC302	43.5443	1580	5	109
TC309	44.008	2011	5	128
TC310	41.836	1427	5	112
TC312	43.131	1252	5	107
TC314	61.4996	1387	5	164
TC316	43.0556	936	5	86
TC318	41.0105	1524	5	120
TC320	46.0208	867	5	100
TC321	39.7751	978	5	88
TC322	42.9142	1002	5	70
TC328	44.9904	1567	5	113
TC333	46.3175	1833	5	106
TC334	41.328	1747	5	109
TC343	47.9592	882	5	126
TC346	45.5232	927	5	201
TC349	42.5039	1294	4	106
TC355	42.7704	823	4	88
TC359	39.0467	986	4	96
TC360	51.9417	1442	4	445
TC364	46.1459	973	2	207
TC365	49.0725	593	2	179
TC366	41.3333	975	4	79
TC367	46.1847	996	4	111
TC368	57.0248	847	4	221
TC369	37.2632	1900	4	84
TC375	39.6064	813	4	119
TC380	44.5274	804	3	109
TC381	42.7981	1451	4	129

© University of Pretoria

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC386	39.8357	1461	3	94
TC390	46.2766	564	2	125
TC392	45.7944	642	4	76
TC393	44.8276	1015	4	85
TC394	43.7728	1919	4	165
TC396	42.8153	682	4	148
TC398	40.8994	934	4	114
TC399	40.9874	871	4	120
TC4	37.515	2495	28	72
TC403	41.6994	1271	3	107
TC405	51.8805	904	4	162
TC406	39.6185	1363	4	130
TC407	42.8324	1730	4	115
TC411	56.9839	809	2	224
TC412	57.5434	749	2	134
TC415	51.2876	932	4	291
TC417	41.8418	999	2	89
TC420	41.5138	872	4	91
TC421	51.2589	1549	4	130
TC425	43.7292	901	4	91
TC434	42.7225	955	4	78
TC435	38.0111	905	4	77
TC439	43.6559	930	4	114
TC441	42.1371	992	4	115
TC442	44.8109	1031	4	173
TC445	53.271	1605	4	226
TC449	44.2417	1589	4	95
TC452	47.5776	1610	4	167
TC453	42.3192	1673	4	98
TC454	51.7038	1526	4	295
TC455	42.0319	1004	4	72
TC456	37.3089	981	4	123
TC458	41.844	846	4	81
TC46	48.4271	1049	16	113
TC462	47.8398	949	4	86
TC465	40.4051	938	3	83
TC469	45.323	1037	4	82
TC470	44.1777	833	2	81
TC474	58.0668	1407	4	177
TC478	57.4444	900	4	296
TC483	53.0369	922	3	119
TC484	43.1333	983	2	73
TC485	43.6631	1294	3	137
TC487	38.0338	1241	3	103

© University of Pretoria

Appendix 1. Table A1.3

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC488	46.3812	981	3	254
TC490	49.5818	837	3	142
TC492	60.1333	750	3	203
TC493	61.0409	1345	3	192
TC496	36.9598	1342	3	124
TC497	49.6173	784	3	114
TC499	37.0918	1286	3	91
TC5	47.6354	2622	49	148
TC50	45.2769	1228	8	178
TC500	39.6702	1031	2	79
TC502	40.0507	789	3	79
TC506	36.9427	942	2	78
TC507	50.7576	924	3	89
TC511	46.8072	1613	3	100
TC518	52.3566	976	2	196
TC526	44.9481	1059	3	108
TC531	43.9306	1211	3	151
TC532	45.4046	1273	2	82
TC536	52.0334	959	2	224
TC538	55.9627	1610	3	178
TC543	38.0471	891	3	51
TC545	44.0318	754	3	119
TC551	42.6392	1311	3	100
TC553	61.6374	855	2	195
TC559	40.8482	1344	3	117
TC56	51.3683	1279	23	200
TC560	47.4537	864	2	156
TC561	47.5198	1008	3	169
TC562	39.372	828	3	81
TC567	43.505	2117	3	108
TC568	40.5267	1367	3	92
TC575	40.4329	1155	3	73
TC577	43.2703	899	3	84
TC580	50.1235	1215	3	390
TC583	35.7697	903	3	90
TC587	45.5844	770	3	197
TC590	38.8337	1612	3	65
TC594	51.595	1442	3	446
TC600	40.8696	920	3	77
TC602	48.4818	988	3	113
TC604	40.2052	2047	3	146
TC605	42.8357	1333	3	99
TC608	46.1957	552	3	96
TC609	60.2808	997	3	309
TC610	47.1698	1272	3	84

© University of Pretoria

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC618	43.5185	864	3	107
TC620	54.8387	899	3	246
TC622	45.2351	787	3	140
TC623	39.1787	901	3	69
TC625	45.7045	582	3	68
TC626	47.9003	1524	3	386
TC627	47.2691	1007	3	119
TC628	42.56	625	3	166
TC631	40.9449	1524	3	102
TC633	60.9058	839	3	200
TC634	53.2972	1107	2	312
TC635	40.5337	787	2	62
TC64	49.5446	2196	13	606
TC642	42.0753	877	3	90
TC644	40.8009	924	3	119
TC647	40.3701	1459	3	106
TC65	48.14	1371	5	202
TC656	41.3965	802	3	88
TC657	41.3502	948	3	94
TC658	40.4624	865	3	115
TC66	50.4263	821	2	247
TC664	51.4486	1001	3	135
TC665	41.478	839	3	61
TC666	42.8424	971	3	98
TC669	60.5701	842	3	135
TC67	63.3374	821	9	199
TC674	45.0435	575	3	96
TC678	50.7502	1533	3	510
TC679	43.8613	1067	3	109
TC68	63.6964	909	6	230
TC680	43.0215	1304	2	89
TC684	39.8453	1034	3	79
TC685	34.6591	880	3	63
TC686	51.8968	659	2	159
TC687	48.1818	660	2	132
TC691	46.3926	1289	3	108
TC693	69.2884	801	2	266
TC696	45.0499	1303	2	123
TC699	51.2451	763	2	99
TC701	40.1891	846	2	125
TC704	40.7013	1597	2	96
TC705	38.7097	806	2	87
TC706	45.0935	856	2	90
TC71	41.8574	1529	20	112

© University of Pretoria

Appendix 1. Table A1.3

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC710	48.4211	1615	2	97
TC713	39.976	833	2	99
TC714	53.0806	1477	2	139
TC715	44.5521	826	2	144
TC720	38.9943	1054	2	68
TC724	50	462	2	111
TC725	44.8892	1399	2	98
TC726	56.5094	1060	2	223
TC728	61.4013	785	2	212
TC730	40.4884	778	2	113
TC731	38.1176	850	2	92
TC733	43.9512	901	2	92
TC734	37.4642	1396	2	109
TC737	42.1818	825	2	75
TC738	39.4422	1255	2	89
TC743	44.4206	932	2	78
TC744	45.3592	1239	2	101
TC746	49.434	795	2	222
TC747	61.6592	446	2	100
TC748	49.0591	1541	2	209
TC750	45.6422	872	2	70
TC751	39.3839	909	2	67
TC752	44.9555	674	2	118
TC753	47.486	895	2	115
TC755	36.2832	904	2	111
TC756	50.7819	1215	2	185
TC763	39.8148	1404	2	117
TC764	50.922	1410	2	151
TC766	44.0079	509	2	153
TC769	39.5639	963	2	72
TC770	37.1186	590	2	95
TC771	50.6122	1715	2	136
TC772	45.8856	717	2	118
TC776	41.8087	763	2	91
TC777	46.6964	1347	2	96
TC78	41.8209	1351	19	135
TC780	48.35	1303	2	271
TC781	47.0164	1525	2	189
TC782	47.027	740	2	193
TC783	48.8055	293	2	56
TC784	43.4266	1430	2	185
TC785	45.4545	1166	2	101
TC787	36.7869	1525	2	90
TC788	41.6599	1229	2	168
TC79	48.7163	1558	17	479

© University of Pretoria

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC793	42.3032	1702	2	129
TC794	43.0493	223	2	46
TC796	52.5632	1385	2	441
TC798	43.8167	1593	2	128
TC802	34.4051	311	2	51
TC806	39.4536	915	2	73
TC810	50.9677	310	2	73
TC812	43.5644	909	2	125
TC814	50.9956	904	2	273
TC816	43.2271	1506	2	281
TC817	57.5793	851	2	176
TC819	36.7781	329	2	87
TC823	40.9149	787	2	66
TC825	42.4142	903	2	108
TC826	43.3933	1279	2	120
TC827	37.5573	1310	2	81
TC829	39.976	833	2	72
TC830	40.2027	888	2	85
TC831	37.2574	1597	2	94
TC834	38.0917	1331	2	67
TC837	42.1635	1627	2	111
TC839	40.6934	548	2	85
TC841	46.6525	941	2	192
TC843	40.9783	920	2	106
TC846	44.4515	1577	2	130
TC850	42.6729	853	2	92
TC851	49.5845	361	2	106
TC852	42.0074	538	2	71
TC853	42.723	852	2	77
TC854	51.4175	776	2	198
TC855	59.1413	722	2	155
TC857	44.4763	697	2	107
TC858	40.669	1704	2	127
TC859	38.9386	1564	2	110
TC861	50.4188	597	2	125
TC862	39.4479	1630	2	79
TC863	41.6887	758	2	84
TC864	49.8039	765	2	109
TC866	51.9897	779	2	90
TC867	61.1364	880	2	239
TC869	44.2927	1025	2	156
TC870	38.8828	913	2	113
TC872	49.3017	1432	2	164
TC880	57.7143	525	2	109

© University of Pretoria

Appendix 1. Table A1.3

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC881	41.4359	975	2	67
TC882	54.7352	623	2	157
TC883	42.1356	1386	2	119
TC884	44.8311	977	2	105
TC885	54.717	689	2	160
TC886	39.1967	722	2	96
TC887	63.1646	790	2	253
TC889	45.183	1038	2	144
TC890	40.7143	1400	2	101
TC891	47.0358	1788	2	129
TC892	41.4758	1179	2	92
TC894	47.046	1371	2	90
TC897	48.8756	1334	2	168
TC898	42.9558	724	2	187
TC899	37.5194	1290	2	95
TC9	46.433	1626	26	129
TC902	41.639	903	2	133
TC905	36.143	783	2	92
TC908	37.1459	953	2	107
TC909	40.0124	1612	2	99
TC91	46.0127	1417	16	117
TC910	39.9142	932	2	79
TC911	25.1989	377	2	106
TC912	44.9097	1385	2	97
TC913	36.3041	1526	2	134
TC914	39.3333	1200	2	106
TC915	49.6215	1189	2	128
TC917	49.2683	820	2	107
TC918	38.7705	1220	2	68
TC919	51.866	1313	2	293
TC920	49.3741	719	2	73
TC921	41.7024	1398	2	83
TC922	39.2821	975	2	96
TC924	35.1201	541	2	57
TC925	43.0493	892	2	123
TC926	47.3723	1351	2	216
TC927	40.3935	864	2	98
TC930	42.3767	892	2	92
TC931	50.3893	899	2	96
TC933	54.7117	1422	2	282
TC934	42.0891	651	2	83
TC935	39.8374	861	2	78
TC938	40.2899	1380	2	104
TC941	59.9732	1494	2	325
TC943	45.8207	1316	2	210

© University of Pretoria

TC	%GC	Length (bp)	Num. ESTs	Longest ORF (aa)
TC944	43.4003	1394	2	121
TC947	41.9847	393	2	114
TC948	44.2623	244	2	77
TC950	41.0412	1652	2	99
TC952	55.168	1577	2	398
TC954	43.2111	1009	2	130
TC957	40.796	804	2	67
TC959	61.175	783	2	178
TC96	39.0918	1013	14	110
TC960	38.3467	871	2	156
TC962	50.431	928	2	184
TC970	39.2781	942	2	89
TC973	57.8834	926	2	116
TC975	40.5896	882	2	85
TC977	57.931	870	2	268
TC980	35.7697	903	2	53
TC982	41.1765	1037	2	85
TC985	64.0223	895	2	180
TC991	39.5667	877	2	87
TC993	51.9452	1388	2	125
TC996	51.9435	849	2	228
TC997	46.5098	1275	2	128
TC998	44.1948	801	2	90

© University of Pretoria

Table A1.4 Significant matches of RaGI-unique sequences (having coding probability > 90%) to known protein domains in the conserved domain database (CDD).

TC	Hit type	PSSM-ID	From	To	E-Value	Bitscore	Accession	Short name	Superfamily	Description	coding probability
TC1024	specific	128862	36	122	2.48E-16	71.2375	smart00595	MADF	cl15779	DNA-binding domains	99.43%
TC1354	non-specific	198389	44	121	6.36E-07	44.2786	cd10442	GIY-YIG_PLEs	cl15257	Nuclease domain found in proteins with cellular functions (also found in transposons)	99.53%
TC997	non-specific	198389	35	62	3.24E-06	41.5822	cd10442	GIY-YIG_PLEs	cl15257	Nuclease domain found in proteins with cellular functions (also found in transposons)	99.37%
TC952	specific	179049	112	136	1.24E-03	36.433	PRK00504	rpmG	cl00383	50S ribosomal protein L33 (role in translation)	98.54%
TC346	non-specific	111039	52	187	2.19E-03	35.8453	pfam02098	His_binding	cl03446	Tick Histamine-binding protein	99.55%
TC488	multi-dom	147578	4	52	3.53E-03	33.2343	pfam05470	eIF-3c_N	-	N-terminus of eukaryotic translation initiation factor 3 subunit	99.64%
TC159	non-specific	184268	42	88	6.23E-03	32.1423	PRK13712	PRK13712	cl10138	conjugal transfer protein TrbA	96.89%
TC1564	multi-dom	179109	18	55	6.51E-03	32.4783	PRK00750	lysK	-	lysyl-tRNA synthetase - catalyses formation of Lys-tRNA	99.25%
TC1951	non-specific	133114	14	79	7.24E-03	31.9275	cd06904	M14_MpaA_like	cl11393	zinc-binding carboxypeptidase	98.97%

© University of Pretoria

Table A1.4

TC	Hit type	PSSM-ID	From	To	E-Value	Bitscore	Accession	Short name	Superfamily	Description	coding probability
TC381	non-specific	192093	34	75	8.18E-03	32.2961	pfam08612	Med20	cl07290	TATA-binding related factor (TRF) of subunit 20 of Mediator complex	99.92%
TC421	multi-dom	171438	6	57	8.44E-03	32.1853	PRK12363	PRK12363	-	phosphoglycerol transferase I	90.03%
TC2242	non-specific	202519	25	49	8.72E-03	31.4304	pfam03055	RPE65	cl10080	Retinal pigment epithelial membrane protein	93.84%
TC2502	non-specific	147978	38	61	8.87E-03	29.4945	pfam06107	DUF951	cl01864	Domain of unknown function. May bind to nucleic acids	97.42%
TC456	non-specific	202436	1	34	9.79E-03	29.9495	pfam02865	STAT_int	cl03749	Signal Transducers and Activators of Transcription (STAT) protein-protein interaction domain	99.62%

© University of Pretoria

Table A1.5 Transposable elements identified by sequence search against Censor database of a) R. appendiculatus genomic contigs and b) RaGI

a)

Repeat Class	**Number of sequences**	**Length of TE segment**
Interspersed Repeat	5	387
DNA transposon	143	12522
Academ	1	55
Chapaev	1	44
Crypton	1	59
EnSpm	33	3299
Harbinger	4	288
Helitron	7	324
Kolobok	1	54
Mariner/Tc1	8	554
Merlin	1	64
MuDR	10	852
piggyBac	1	175
P	4	456
Polinton	5	337
Sola	4	899
Transib	1	234
hAT	31	2421
Endogenous Retrovirus	27	1850
ERV1	10	797
ERV2	13	824
ERV3	3	182
LTR Retrotransposon	182	24068
BEL	14	772
Copia	22	1485
DIRS	2	126
Gypsy	138	21256
Non-LTR Retrotransposon	90	7643
CR1	6	886
Crack	2	198
Daphne	4	253
I	2	184
Jockey	3	179
Kiri	1	63
L1	12	684
L2	5	403
L2B	3	182
NeSL	1	66

© University of Pretoria

Table A1.5

Repeat Class	**Number of sequences**	**Length of TE segment**
Outcast	1	68
Penelope	5	719
Proto1	1	55
Proto2	1	48
R1	4	213
R4	1	53
RTE	6	1280
SINE	24	1622
SINE2/tRNA	12	774
Tad1	3	231
Tx1	2	95
Pseudogene	33	4035
rRNA	11	2748
tRNA	22	1287
Repetitive Element	4	899
Simple Repeat	3	333
Satellite	3	333
SAT	1	54
Transposable Element	445	46422
Total	486	51177

b)

Repeat Class	**Number of sequences**
PEN Interspersed Repeat	34
DNA transposon	101
LTR Retrotransposon	243
Non-LTR Retrotransposon	241
Pseudogene	26
Total	645

© University of Pretoria

Figure A2.1

Appendix 2 Figures

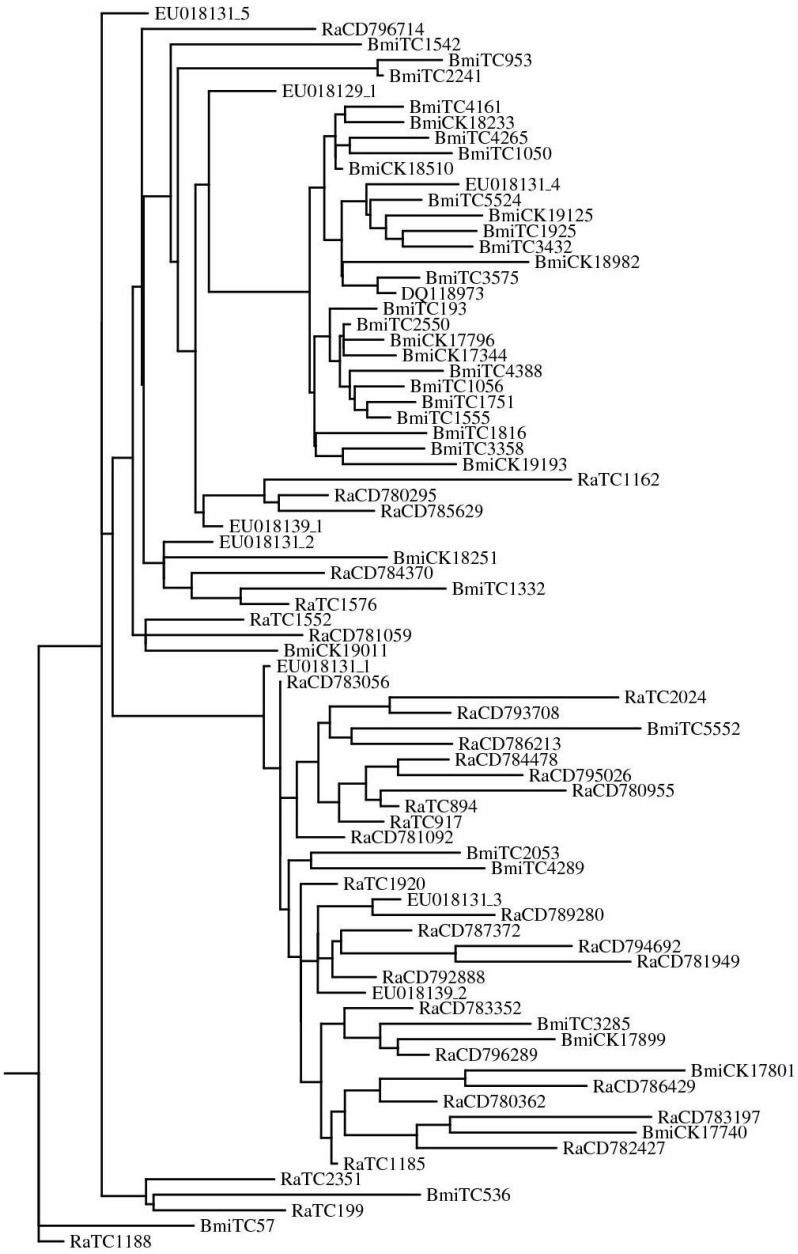


Figure A2.1 Unrooted cladogram illustrating the relatedness of 80 Ruka-like sequences within R. microplus (prefix Bmi) and R. appendiculatus (prefix Ra) transcripts and R. appendiculatus genomic sequences derived from genomic BAC clones (Accession numbers: EU018129, EU018131, EU018139). The cladogram was generated using a maximum likelihood algorithm after calculating 100 bootstrap trees by parsimony.

© University of Pretoria

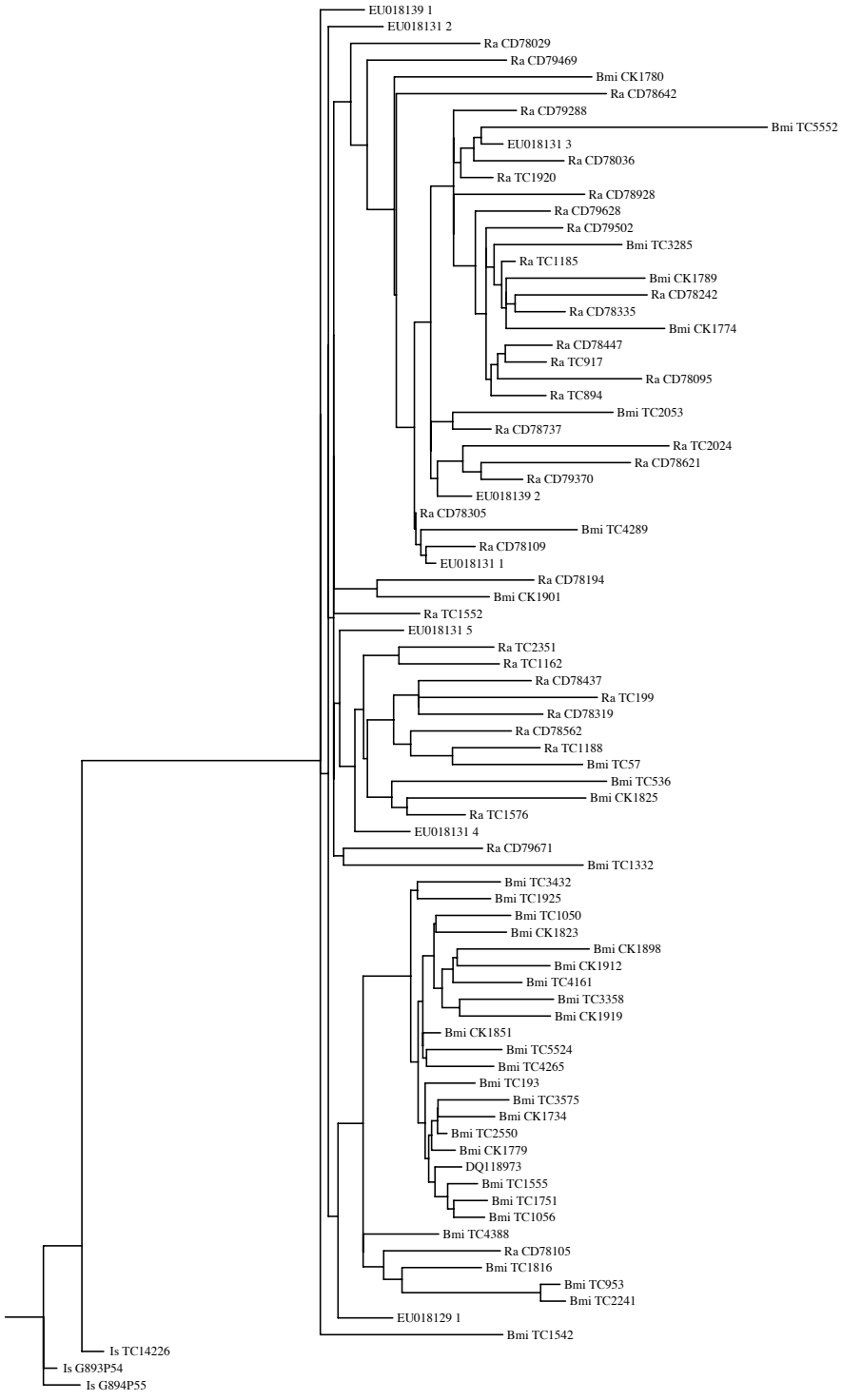


Figure A2.2 Rooted cladogram showing phylogenetic relationship of Ruka-like sequences from R. appendiculatus, R. microplus and I. scapularis (outgroup). Phylogeny computed by the maximum likelihood method.

© University of Pretoria

Figure A2.3

```
CLUSTAL W (1.83) multiple sequence alignment


RAHD87_1        -CCGCCGCGGTGGTATAGCGGTTAGGGTGCTCGGCTGCTGACCCGAAGGTCGCGGGTTC- 58
RAHF80_2        -CCGCCGCGGTGGTGTAGCGGTTACGGCGCTCGGCTGCTGACCCGAAGGTCGCGGGTTC- 58
RAHD87_3        ---GCCGCGGTGGAGCAGTGGTTACGGTGCTCGGTTTCTGACCCGAAGGTTGCGGGTTC- 56
RAHD87_4        -----------------------------CTCGACTGCTGACCCGCAGGTCGCGGGATCT 31
RAHD48_1        ----------TGGTCTAGTGGTTATGGTGCTCGGCTGCTGACCCGCAGGTAGCGGGATCG 50
RAHF80_1        CCCGCTGCGGTGGTCTAGTGGTTACGGGGCTCGACTGCTGACCGGCAGGTCGCGGGTTCG 60
RAHD87_2        CCCGCCACGGTGACCTGGTGGTTATGGTGCTCGACTGCTGACCCGCAGGTCGCGTGATGG 60
RAHD87_5        --------TGTGGACTAGTGGTTATGGTGGTCGACTGCTGACCGGAAGGTTGCGAGATAA 52
                                         ***  * ****** * **** *** * *


RAHD87_1        GATCCCGGCCGCGGCGGTCGCATTT-CGATGGAGGCGAAA-TGGTAGAGGCCCGTGTACT 116
RAHF80_2        AATCCCGGCCGCGGAAGTCGCATTT-CGGTGGAGGCGAAA-TGGTAGAGGCCCGTGTACT 116
RAHD87_3        AATCCCGGCCACGGCGGTCACATTT-CGATGGAGGCGAAA-TGCTTGAGGCCCGTGTACT 114
RAHD87_4        AATCCCGGCACGGGCGGTCGCATTTTCGATGGAGGCGAAAGTGCTTGAGGCTCGTGTGCC 91
RAHD48_1        AATCCCGGCCGCGGCGGCCGCATTTTCTAGGGAGGCGGAAATGCTCGAGGCCCGTGTAC- 109
RAHF80_1        AATCCTTGCCGCGGCGGCCGCATTTTCGATGGAGGCGAAAATGTTTGAGGCCCGTGTGC- 119
RAHD87_2        AATCCCGGCCGTGGTGGCCGCATTTTCGATCGAGGCGAAAATGCTCGAGGCCCGTGTAC- 119
RAHD87_5        AATCCCGGCATCGGCGGCCGCATTT-CGATGAAGGCGAAA-TGCTAGACGCCCGTGTAC- 109
                 ****  **   **  * * ***** *     ***** ** ** * ** ** ***** *


RAHD87_1        GTGCGATGTCAGTGCACGTTAAAGAACACCAGATGGTCGAAATTTCCGGAGCCCTCCACT 176
RAHF80_2        GTGCGATGTCAGTGCACGTTGAAGAACACAAGATGGTCGAAATTTCCGGAGCCCTCCACT 176
RAHD87_3        GTGCGATGTCGGTGCACGTTAAAGAACACGAGATGGTCGAAATTTCCGGAGCCCTCCACT 174
RAHD87_4        TATAGATTTAAGTACACGTTAAAGAACCCCAGGTGGTCAAAATTTTCGGAGCCCTCCACT 151
RAHD48_1        -GTAAATTTAGGTGCACATTAAAGAACCCCAGGTGGTTGAAACGCCCCGAGCCTTCCACT 168
RAHF80_1        -TTAGATTTAGGTGCACGTTAAAGAACCCCAGGTGGTCGAAATTTCCGGAGCCCTCCACT 178
RAHD87_2        -TTAGATATAGGTGCACGTTAAAGAATCCCAGGTGGTCTAAATTTCCGCAGCCCTCCACT 178
RAHD87_5        -TTAGATTTAGGTGCACGCTAAAGAACCCCAGGTGCTCGAAATTTCCGGAGCCCTCCACT 168
                     ** *  ** ***  * *****  * ** ** *  ***    *  **** ******


RAHD87_1        A--CGGCGTCTCTCATAATCATGGTGTGGTTTTGGGACGTTAAACCCCAGATAT 228
RAHF80_2        A--CAGCGTCTCCCATAATCATACCGTGGTTTTGGGACGTAAAACACCAGATAT 228
RAHD87_3        A--CGGCGTCTCTCATAATCGTATCGTGGTTTTGGGACGTAAAACCCCA----- 221
RAHD87_4        A--CGGCGTCTCTCATAACCATATCGTGGTTTTGGGACGTTAAACCCCA----- 198
RAHD48_1        G--CGGCGTCTCTCATAATCATATGGTGGTTTTGGGACGTTAAACCCCA----- 215
RAHF80_1        ATACGGCGTACCTCATAATGATATCGTGGTTTTTGGACGTTAGACCCCA----- 227
RAHD87_2        A--CGGCGTCTCTCATAATCAT---------------------------------- 198
RAHD87_5        A--CGGCGTCCCTTATAATCAT---------------------------------- 188
                   * ****  *  ****   *
```

Figure A2.3 Sequence alignment of eight conserved Ruka copies identified in RA BAC sequences. Primers for the amplification of Ruka copies in R. appendiculatus genomic DNA were designed on this alignment and are underlined. Three forward primer and two reverse primers were selected.

© University of Pretoria