UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Ehrlichia ruminantium: Genome assembly and analysis with the identification and testing of vaccine candidate genes

by

JUNITA LIEBENBERG

Submitted in partial fulfilment of the requirements for the degree Doctor of Philosophy

in the Department of Veterinary Tropical Diseases, Faculty of Veterinary Science,

University of Pretoria

September 2010

ACKNOWLEDGEMENTS

I would like to express my gratitude to my promoters Prof. Basil Allsopp and Dr. Nicola Collins for the opportunity to obtain a Ph.D. at the University of Pretoria. Special thanks to Prof. Allsopp for his mentorship and for the time he spent in correcting my writing. My thanks also go to Dr. Collins for her valuable advice, scientific contribution and motivation.

I would also like to extend my appreciation to my colleagues at the ARC-OVI for their contributions in the animal trials and immunological assays. A special word of thanks Dr. Mirinda van Kleef and Dr. Alri Pretorius for their assistance with the lymphocyte proliferation and ELISpot assays, and their expertise in cellular immunology. I would also like to thank Dr. Erich Zweygarth and Antoinette Josemans for providing the *E. ruminantium* cell cultures and Helena Steyn for her assistance with the immunisation of animals.

The work presented in this thesis was supported by the Department of Science and Technology of South Africa (LEAD 37/2001 (87)), the European Union (FP6-003713), the National Research Foundation of South Africa (FA2004042200063) and the Agricultural Research Council of South Africa.

Finally special thanks to Frans for his understanding and support.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

A	adenine
aa	amino acids
ABC	ATP-binding cassette
ACT	Artemis Comparison Tool
ADP	adenosine diphosphate
ATP	adenosine triphosphate
BAC	bacterial artificial chromosome
BCG	bacillus Calmette Guérin
bp	base pairs
BSA	bovine serum albumin
C	cytosine
CD	cluster of differentiation
CDS	coding sequence
cfu	colony forming units
CoA	coenzyme A
ConA	Concanavalin A
cpm	counts per minute
CTL	cytotoxic T-lymphocytes
DHF	dihydrofolate
dNTP	deoxynucleotide tri-phosphate
DNA	deoxyribonucleic acid
EC	Enzyme Commission
EDTA	ethylene diamine tetraacetic acid
ELISA	enzyme-linked immunosorbent assay
ELISpot	enzyme-linked immunosorbent spot
FACS	fluorescent-activated cell sorting
FAD	flavin adenine dinucleotide
FCS	foetal calf serum
G	guanine
Gb	gigabase(s)
His	histidine
HRP	horseradish peroxide
IFA	indirect fluorescent antibody
IFN-γ	interferon-gamma

IgG	immunoglobulin G
IHF	integration host factor
IL	interleukin
IPTG	isopropyl-β-D-thiogalactoside
kb	kilobase(s)
kDa	kilodalton
kPa	kilopascal
LB	Luria-Bertani
LD_{50}	lethal dose, 50%
LPA	lymphocyte proliferation assay(s)
LTRs	longer tandem repeats
µCi	microcurie
M	molar
MAP	major antigenic protein
Mb	megabases
MMR	measles, mumps and rubella
mRNA	messenger ribonucleic acid
MW	molecular weight
NAD^+	nicotinamide adenine dinucleotide
NADH	nicotinamide adenine dinucleotide - hydrogen (reduced)
NK	natural killer
N-terminal	amino terminal
OMP	outer membrane protein
ORF	open reading frame
PBMC	peripheral blood mononuclear cells
PBS	phosphate buffered saline
PBS-T	phosphate buffered saline-Tween
PCR	polymerase chain reaction
PT	pertussis toxin
PVDF	polyvinylidene fluoride
r	recombinant
RBS	ribosomal binding site
RI	reaction index
RNA	ribonucleic acid
RNA-Seq	ribonucleic acid sequencing
rpt	repeat
rRNA	ribosomal ribonucleic acid

ru	repeat unit
SDS	sodium dodecyl sulphate
SFC	spot forming cells
SI	stimulation index
SPG	sucrose potassium glutamate
ssDNA	single-stranded DNA
SSRs	simple sequence repeats
T	thymine
TCA	tricarboxylic acid
th	transmembrane helix
Th1	T-helper 1
tmRNA	transfer-messenger ribonucleic acid
TNF-β	tumour necrosis factor-beta
tRNA	transfer ribonucleic acid
U	enzyme unit(s)
Vlp	variable surface lipoprotein

SUMMARY

Ehrlichia ruminantium: Genome assembly and analysis, with the identification and testing of vaccine candidate genes

by

JUNITA LIEBENBERG

PROMOTOR: Prof. B. A. Allsopp

CO-PROMOTOR: Dr. N. E. Collins

DEPARTMENT: Veterinary Tropical Diseases

DEGREE: Ph.D.

A shotgun genome sequencing project was undertaken in the expectation that access to the entire protein coding potential of *E. ruminantium* (Welgevonden) will facilitate the identification of vaccine candidate genes against heartwater. The 1,516,355 bp sequence is predicted to encode 888 proteins and 41 stable RNA species. The most prominent feature is the large number of tandemly repeated and duplicated sequences, some of continuously variable copy number. These repeats have mediated numerous translocation and inversion events and seem to be responsible for the generation of both new full and partial protein coding sequences. There are 32 predicted pseudogenes, most of which are truncated fragments of genes associated with repeats. Of the 13 members of the order Rickettsiales compared in this study, *E. ruminantium* has the lowest coding capacity (62%), lowest GC content (27.5%), but the highest proportion of repetitive sequences, which comprise 8.5% of the genome. Metabolic reconstruction of *E. ruminantium* revealed the metabolic and biosynthetic capabilities typical of an obligate intracellular organism. We identified a number of genes unique to *E. ruminantium,* most of which are not functionally

characterised in any organism, and those shared with 12 other members of the Rickettsiales. Bioinformatic tools were used to identify possible vaccine candidates from the annotated genome sequence. The protective properties of seven open reading frames (ORFs), which induced cellular immune responses *in vitro*, were tested *in vivo*. Only 20% survival was obtained in sheep immunised with a DNA formulation consisting of three ORFs. We found that the levels of peripheral blood mononuclear cell proliferation and interferon-gamma (IFN-γ) production did not correlate with each other, nor with the levels of protection, suggesting that the current assays are just not reliable and that IFN-γ expression alone is not an indicator of protection. Therefore more cytokines and different assays will have to be investigated to define in detail what constitutes a protective immune response against *E. ruminantium* infection. However, the data generated from the genome sequence will continue to facilitate novel approaches to study the organism and to develop an efficacious vaccine against heartwater.

CHAPTER 1

Literature review

1.1. HEARTWATER

1.1.1. History of heartwater research

In 1838 the Voortrekker pioneer Louis Trichardt documented a fatal disease amongst his sheep, following a massive tick infestation (Neitz, 1968). This is believed to be the first record of heartwater, a tick-borne disease affecting wild and domestic ruminants throughout sub-Saharan Africa, including the islands of Zanzibar, Mauritius, Madagascar, Sao Tomé and Réunion (Uilenberg, 1983; Provost & Bezuidenhout, 1987; Flach *et al.*, 1990), and the French Antilles (Muller Kobold *et al.*, 1992; Camus & Barré, 1995).

In 1898 it was shown that the disease could be transferred from diseased to susceptible animals by blood passage (Dixon, 1898; Edington, 1898) and Hutcheon (1900) concluded that heartwater was caused by a living microorganism. At first it was thought that the disease-causing organism was a virus (Spreull, 1904), but in 1925 Cowdry demonstrated that heartwater was caused by an intracellular rickettsial bacterium, which he called *Rickettsia ruminantium* (Cowdry, 1925a, b). Later the name was changed to *Cowdria ruminantium* (Moshkovski, 1947), and recently the organism was reclassified as *Ehrlichia ruminantium* (Dumler *et al*., 2001).

The first effective - and still the only - commercially available method of immunization, the so-called blood vaccine, was introduced by Neitz and Alexander in the 1940s (Neitz & Alexander, 1941, 1945; Oberem & Bezuidenhout, 1987). Another significant development with regard to the control of heartwater was the discovery of effective curative drugs, such as sulphonamides and tetracyclines (Neitz, 1940; Weiss *et al.*, 1952; Haig *et al.,* 1954).

The discovery of an isolate that is highly pathogenic to mice (Du Plessis & Kümm, 1971) facilitated the development of a mouse model for heartwater research, and the successful *in vitro* cultivation of the organism in 1985 (Bezuidenhout *et al.,* 1985) has enabled researchers to produce large quantities of the organisms to study at the molecular level. Recently the *in vitro* culture system has been improved with the use of chemically defined media (Zweygarth & Josemans, 2001a) and by the propagation of *E. ruminantium* in tick cell lines (reviewed by Bell-Sakyi *et al*., 2007).

1.1.2. The organism

E. ruminantium is a Gram-negative, α-proteobacterium, belonging to the family Anaplasmataceae, order Rickettsiales. All organisms in the order Rickettsiales are obligate intracellular bacteria, but members of the family Anaplasmataceae are found within membrane-bound vacuoles whereas members of the family Rickettsiaceae grow freely within the cytoplasm of eukaryotic cells. The genus *Ehrlichia* also includes the canine and human pathogens *E. canis, E. ewingii* and *E. chaffeensis* (Dumler *et al*., 2001).

Ticks acquire the bacteria while feeding on an infected host. In the tick gut cells the organisms multiply and then spread to the haemolymph and salivary glands (Kocan & Bezuidenhout, 1987). *E. ruminantium* is transmitted through the saliva to the vertebrate host (Kocan *et al*., 1987) and it primarily infects vascular endothelial cells (Cowdry, 1926), however some strains have also been observed in circulating leukocytes (Logan *et al*., 1987). In the host cells the organisms are enclosed in a vacuole surrounded by a membrane derived from the host cell membrane, here they replicate mainly by binary fission to form large colonies of metabolically active reticulate bodies (Prozesky & Du Plessis, 1987). Five to six days after infection the cell disrupts to release infectious electron-dense elementary bodies.

The traditional microscopical detection of the organisms in Giemsa/Diff-Quick stained brain smears is still the most commonly used method to confirm that an animal has died of heartwater (Camus & Barré, 1987). A range of serological tests (indirect fluorescent antibody (IFA), enzyme-linked immunosorbent assay (ELISA) and Western blots) are available, but they are compromised by cross-reacting with other *Ehrlichia* spp. (Du Plessis *et al*., 1993). DNA-based tests have been developed which are more sensitive and specific than the serological assays; the new tests use *E. ruminantium* targets such as pCS20 (Waghela *et al*., 1991; Van Heerden *et al*., 2004b; Steyn *et al*., 2008), *map*1 (Kock *et al*., 1995) and the 16S rRNA gene (Allsopp *et al*., 1997).

1.1.3. The vector

Twelve species of *Amblyomma* ticks are known to be capable of transmitting the disease; two of these, *A. variegatum* and *A. hebraeum,* are of major importance in Africa (Walker & Olwage, 1987). The only vector in South Africa is *A. hebraeum* and this was the first vector of the disease to be identified (Lounsbury, 1900). *A. variegatum* is the most widely distributed vector in Africa and it is also well established on many islands in the Caribbean Sea. Heartwater, however, is only established on three islands in the Lesser Antilles to which infected ticks were probably originally introduced from Africa (Uilenberg, 1990). These infected ticks could have been introduced into Guadeloupe during the nineteenth century with cattle from Senegal (Curasson, 1943), although it is also possible that the introduction was as early as the eighteenth century (Maillard & Maillard, 1998). From the Caribbean region heartwater poses the threat of spreading to the American mainland, where *A. maculatum* and the white tailed deer (*Odocoileus virginianus*) already constitute a viable native tick-host pair for the maintenance of *E. ruminantium* (Uilenberg, 1982; Barré *et al*., 1987; Mahan *et al*., 2000).

Amblyomma ticks infest cattle, sheep, goats, horses and wild game, including reptiles, birds and mammals (reviewed by Allsopp *et al*., 2004). Adults usually attach on the underside of the body,

while nymphs are mostly recovered from the feet, and larvae can be found on the head and feet (reviewed by Petney *et al*., 1987). Three hosts are required to complete the life cycle of *Amblyomma* ticks, since the larvae and nymphs need to feed on a host before they drop off to moult. All three life cycle stages, larvae, nymphs and adults, can become infected, and larvae and nymphs subsequently become infective at the following instar (reviewed by Bezuidenhout, 1987).

1.1.4. The disease

Heartwater is considered to be one of the most important endemic diseases of domestic livestock in southern Africa. Economic losses occur as a result of high mortality rates, which can be up to 90% in susceptible animals (Neitz, 1964; Du Plessis & Malan, 1987), the costs of control, as well as restrictions being placed on the export of animals and animal products. The disease is a major problem when susceptible animals are moved from heartwater-free to heartwater-infected areas (Neitz, 1968; Simpson *et al*., 1987) and is a significant obstacle to the introduction of high-producing animals to upgrade local stock (Kanyari & Kagira, 2000).

The incubation period and the severity of the disease depend on the age and breed of the animal affected and the virulence of the heartwater isolate. It usually takes less than two weeks for the disease to manifest and early clinical signs include an elevated temperature, often exceeding 41^{o}C, respiratory distress, loss of appetite and diarrhoea. This is often followed by nervous symptoms, such as constant movement of the lower jaw and tongue, incoordination, muscular twitching and squinting. The onset of nervous symptoms is usually followed by death within 48 h. Accumulation of fluid in the chest cavity is common in most fatal cases of the disease and the name heartwater is derived from the presence of fluid in the heart sac. Fluid in the lungs often coagulates on exposure to air, which leads to a frothy discharge from the nostrils and mouth (reviewed by Van de Pypekamp & Prozesky, 1987).

1.1.5. Immune responses to *E. ruminantium* infection

It is generally accepted that cellular immunity plays an important role in host defence against intracellular bacteria. Cell-mediated immunity involves several mechanisms: the activation of antigen-specific cytotoxic T-lymphocytes (CTLs) that are able to lyse body cells which display epitopes of foreign antigen on their surface; the activation of macrophages and natural killer (NK) cells, enabling them to destroy intracellular pathogens; and the secretion of a variety of cytokines (Roitt, 1991). Cytokines influence the activity of a variety of body defence cells as well as stimulate various non-specific body defences such as inflammation and fever.

T-cell responses characterised by $CD4^+$, $CD8^+$ and γδ T-cells, combined with the expression of interferon-gamma (IFN-γ), IFN-alpha, tumour necrosis factor-beta (TNF-β) and interleukin-2 (IL-2), have all been implicated in protective immunity to heartwater (Du Plessis *et al*., 1991, Totté *et al*., 1997; Mwangi *et al*., 1998; Byrom *et al*., 2000). The cytokines IFN-γ, TNF-β and IL-2 are produced by T helper (Th) 1-lymphocytes upon the recognition of antigens presented by macrophages. These cytokines collectively enable $CD8^+$ T-cells to proliferate and differentiate into CTLs capable of destroying infected host cells, and also activate NK cells, macrophages and neutrophils (Mosmann & Sad, 1996; Ojcius *et al*., 1996; Harding *et al*., 2003; Chabalgoity *et al*., 2007).

IFN-γ has been shown to be a powerful inhibitor of *E. ruminantium* growth *in vitro* (Totté *et al*., 1993, 1996) and has also been implicated in protection against several other tick-borne diseases of ruminants (Kodama *et al*., 1987; Preston *et al*., 1992; Brown *et al*., 1996, 1999). Therefore, antigens that induce strong cell-mediated immune responses characterised by IFN-γ expression would probably be useful vaccine candidates. Thus far, only two recombinant proteins, major antigenic proteins 1 and 2 (MAP1 and MAP2), have been shown to induce T-cell lines to produce IFN-γ (Mwangi *et al*., 2002). It has also been found that *E. ruminantium* proteins in the molecular

weight ranges 13-18 kDa (Van Kleef *et al*., 2002) and 22-32 kDa (Esteves *et al*., 2004) induce IFN-γ production, but the specific antigens responsible for this effect have not been identified.

1.1.6. Heartwater vaccine development

Heartwater is routinely controlled by extensive dipping against ticks. However, this strategy is expensive and labour intensive, and ticks often develop resistance against acaricides. The only commercially available immunisation procedure is the infection-and-treatment method developed at Onderstepoort (Oberem & Bezuidenhout, 1987). Animals are infected with sheep blood containing live virulent organisms of the Ball3 isolate, followed by tetracycline treatment during the febrile reaction. Although the infection-and-treatment method has been used with a degree of success it has numerous disadvantages. The frozen blood has to be stored in liquid nitrogen or dry ice up to the time of inoculation, it has to be administered intravenously, and the animals have to be monitored clinically for a febrile reaction, whereupon they must be treated with antibiotics. Furthermore, animals may die as a result of the *E. ruminantium* infection, or of infection with other disease-causing organisms which may be accidentally transmitted with the infected blood. Because of these deficiencies a better vaccine is badly needed and several alternative types are being investigated, including live attenuated, inactivated and DNA vaccines.

1.1.6.1. Attenuated heartwater vaccines

Live attenuated vaccines are usually very effective because they induce both cellular and humoral responses. The first heartwater attenuated vaccine, consisting of tissue culture-derived attenuated organisms, was described for the Senegal isolate (Jongejan, 1991). Although this vaccine conferred protection against homologous challenge it did not provide efficient protection against several other stocks. More recently, attenuation has also been achieved for the Welgevonden isolate (Zweygarth & Josemans, 2001b). Both sheep and goats were protected against a lethal needle challenge with the homologous stock, and it was also shown that sheep were fully protected against four other virulent stocks (Zweygarth *et al*., 2005). Previous studies have also

shown that immunity to the Welgevonden isolate confers immunity to a number of heterologous virulent stocks (Du Plessis *et al*., 1989, Collins *et al*., 2003). Although the Welgevonden attenuated vaccine shows a lot of promise, it still needs to be evaluated in field conditions and the possibility of reversion to virulence limits its use to heartwater-endemic areas.

1.1.6.2. Inactivated heartwater vaccines

Killed inactivated vaccines are safer to use than live attenuated vaccines, yet they may contain undesirable components like bacterial endotoxins, and the presence of numerous non-protective components may reduce the degree of protection achieved. Also, although immunisation with killed organisms induces strong antibody responses, cellular immune responses are typically poor (Dunham, 2002). An inactivated vaccine containing chemically inactivated *E. ruminantium* elementary bodies has been investigated and varying levels of protection were obtained during laboratory trials in goats (Martinez *et al*., 1994), sheep (Mahan *et al*., 1995) and cattle (Totté *et al*., 1997). Although the inactivated vaccine reduced mortality in field trials across southern Africa (Mahan *et al*., 2001), complete protection has not been shown. This vaccine is also difficult and expensive to produce, since large quantities of endothelial cells are required for its preparation (Totté *et al*., 1997).

1.1.6.3. DNA vaccines

Typically, a DNA vaccine consists of the specific gene(s) of interest cloned into a bacterial plasmid engineered for optimal expression in eukaryotic cells. DNA-based vaccines offer a number of advantages over conventional vaccines, such as ease of construction, heat stability, low production cost, an ability to induce strong, polarised Th1-type $CD4^+$ and $CD8^+$ responses, and the ability to produce vaccines for organisms that are difficult or dangerous to culture (reviewed by Huygen, 2003). These vaccines are also believed to be genetically safe, as it has been shown that the risk for homologous recombination and mutagenic integration into the host DNA is very low (Nichols *et al*., 1995). DNA-based vaccination (also known as genetic immunisation) consists of the direct transfer of a naked bacterial plasmid DNA into the animal cells (Davis *et al*.,

1994) where the recombinant pathogen gene(s) are expressed. The products are recognised as foreign and stimulate a protective response from the host immune system.

DNA immunisation has been reported to induce protective immunity against several bacterial pathogens, including *Brucella melitensis* (Yang *et al*., 2005), *B. abortus* (Luo *et al*., 2006), *Chlamydophila abortus*, (Stemke-Hale *et al*., 2005) and *Listeria monocytogenes* (Rapp & Kaufmann, 2004). Moreover, two DNA vaccine products in the area of veterinary medicine have been approved in 2005 (Ulmer *et al*., 2006), one against the West Nile virus in horses (Powell, 2004) and one against infectious haematopoietic necrosis virus in salmon (Lorenzen & LaPatra, 2005).

A DNA vaccine encoding the immunodominant MAP1 protein of *E. ruminantium* was shown to partially protect mice against homologous lethal challenge (Nyika *et al*., 1998, 2002), yet there has been no report of this vaccine protecting ruminants against heartwater. Previous research carried out in our laboratory has identified a cocktail of four *E. ruminantium* open reading frames (ORFs) that induce 100% protection in sheep against a lethal needle challenge when delivered as a DNA vaccine (Collins *et al*., 2003; Pretorius *et al*., 2007). However the same cocktail only induced 20% protection against heartwater in a natural field challenge situation (Pretorius *et al*., 2008).

1.2. GENOME SEQUENCING

1.2.1. DNA sequencing

In 1977 Sanger published his method for determining the order of nucleotide bases of DNA using chain-terminating nucleotide analogues, or dideoxynucleotides (Sanger *et al*., 1977). The same year Maxam and Gilbert (1977) reported their chemical sequencing method, but due to its technical complexity and extensive use of hazardous chemicals it never became as popular as the Sanger method. Strauss and co-workers (1986) improved the Sanger sequencing method by

attaching fluorescent dyes to the dideoxynucleotides, which permitted them to be sequentially detected and read into a computer. A year later Applied Biosystems (http://appliedbiosystems.com) developed the first automated slab gel DNA sequencer, which read fragments as they were separated on a polyacrylamide gel. The slab gels were later replaced by capillaries filled with an electrophoresis medium, which simplified the separation step and increased the length of reads (Madabhushi, 1998). Over the last decade the average length of a sequencing read has increased from approximately 450 bp to 850 bp (Hall, 2007).

In 1995 the first complete genome sequence of a free-living organism was reported, that of *Haemophilus influenzae* (Fleischmann *et al.*, 1995). The genome sequences of hundreds of bacteria and several eukaryotes have subsequently been determined; the organisms include the nematode worm, *Caenorhabditis elegans* (*C. elegans* Sequencing Consortium, 1998); the fruit fly, *Drosophila melanogaster* (Adams *et al*., 2000); the first plant, *Arabidopsis thaliana* (*Arabidopsis* Genome Initiative, 2000); and the human genome (International Human Genome Sequencing Consortium, 2001; Venter *et al.*, 2001). Although these have been major achievements it is only the beginning of a period in which large amounts of sequence information will be required from many individuals and species. The knowledge of multiple genome sequences is an essential tool to investigate complex disease, pathogenicity, evolution and individuality.

For almost 30 years the Sanger sequencing method has been used for DNA sequencing, but more widespread application of conventional sequencing technology is limited by cost, speed, and sensitivity. Hence there is a need for cheap high-throughput sequencing methods that could improve productivity by several orders of magnitude without the need for extensive infrastructure (Bentley, 2006).

1.2.1.1. Novel sequencing technologies

Novel sequencing technologies (also referred to as next-generation, high-throughput, ultra-deep or massively parallel sequencing) can be classified into three main strategies: *in vitro* cloning,

amplification and mass spectrometry, and single-molecule approaches (reviewed by Bentley, 2006; Hall, 2007; Turner *et al*., 2009; and by Voelkerding *et al*., 2009). Although mass spectrometric methods, such as the MassArray method (Jurinke *et al*., 2002), are commonly used for single nucleotide polymorphism analysis these are still very specialised techniques which are not widely used for *de novo* sequencing. The *in vitro* cloning and single molecule approaches will be discussed further.

1.2.1.1.1. In vitro *amplification sequencing technologies*

Since the measuring of biochemical processes at single-molecule resolution is technically demanding, amplification of the DNA is generally employed before sequencing. This is usually done by cloning the DNA into a plasmid and growing clones. However, this approach has its drawbacks, such as the presence of stretches of DNA having physical properties that prevent efficient replication in *E. coli*, or the presence of genes coding for toxic compounds which kill the host cell. Several inexpensive high-throughput strategies for *in vitro* amplification have been developed recently which avoid some of the inherent biases of *in vivo* methods. These technologies are of two types, sequencing by synthesis, for example the 454 (Margulies *et al*., 2005) and Solexa (now Illumina) (Bennett *et al*., 2005) methods, and hybridisation and ligation of oligonucleotides, such as the polony or sequencing by ligation method (Shendure *et al*., 2005).

The 454 system involves massively parallel sequencing by synthesis. In this approach the ssDNA template chains are immobilised and amplified on beads which are individually isolated in the aqueous phase droplets of a reaction buffer in oil emulsion. The beads are then applied to a picotiter plate, in which most cells contain a single bead, for pyroseqeuncing (Ronaghi, 2001). Single deoxynucleoside triphosphate solutions flow sequencially across the picotiter plate one at a time and the polymerase extends the existing DNA strand by adding complementary nucleotide(s). Each base incorporation is detected by the release of a chemiluminescent signal. The 454 technology was taken up by Roche (http://www.roche.com) which introduced it as the Genome Sequencer 20 (GS 20) System in 2005. The Genome Sequencer FLX (Droege & Hill,

2008) was revealed in 2007 and the current system, the Genome Sequencer FLX with GS FLX Titanium series reagents, generates 400 million high quality bases per run at read lengths of approximately 400 bases.

The polony method uses ssDNA template bound to beads, in a similar manner to the 454 method, and the beads are immobilized in a monolayer in an acrylamide matrix for amplification to form polymerase colonies or "polonies". Sequencing is performed using multiple cycles of ligation of fluorescently labeled degenerate nanomers. Only complimentary nanomers will anneal to the anchor primer, which ensures great accuracy. Applied Biosystems (http://appliedbiosystems.com) acquired the polony method and launched it as the SOLiD System in 2007. Currently the SOLiD 3 Plus System generates 25-50 bp reads and yields 60 Gb data per run.

In contrast to Roche 454 or SOLiD, the Solexa system (http://www.solexa.com) amplifies the DNA on a solid surface. Sequencing by synthesis is carried out by incorporating modified nucleotides linked to coloured dyes and the presence of all four bases in the reaction mixture minimises the risk of misincorporation. The Solexa Genome Analyzer, the first "short read" sequencer released in 2006, generates reads of 18-35 bases to yield up to 1 Gb per run. The technology is now incorporated in the Illumina Genome Analyzer System (http://www.illumina.com) and produces 5-33 Gb data per run of 35-100 bp reads.

1.2.1.1.2. Single-molecule sequencing technologies

The single-molecule approaches are still in the developmental stages and are therefore sometimes referred to as the third generation or even "next-next" generation of sequencing platforms. These techniques can overcome many of the problems that result from the amplification of DNA which is needed by other technologies. Braslavsky and co-workers (2003) reported the use of DNA polymerase to obtain sequence information from single DNA molecules by using fluorescence microscopy. Fluorescently labelled nucleotides are incorporated into individual DNA strands with single base resolution. This method generates fingerprints up to 5 bp in length only, but

since the technology has been commercialised by Helicos Biosciences Corporation longer reads (25 to 50 bp) have been reported (http://www.helicosbio.com; Voelkerding *et al*., 2009). The Helicos platform HeliScope, launched in 2008, was used successfully to resequence the 6,407 bp genome of bacteriophage M13 (Harris *et al*., 2008).

Another method of single-molecule sequencing involves "reading" the physical properties of the DNA molecule as it is passed through a nanopore (Kasianowicz *et al*., 1996; Storm *et al*., 2005). In theory, this method offers unlimited read lengths and once the technical difficulties are overcome it could revolutionise genome sequencing. For example, Oxford Nanopore Technologies is developing a label-free single-molecule sequencing technology using an α-haemolysin nanopore. This method and several other emerging single-molecule sequencing approaches are reviewed by Ansorge (2009) and by Turner and colleagues (2009).

1.2.1.1.3. Limitations and advantages of the latest technologies

The major difficulty with all of the next-generation technologies is that the very short reads present a challenge in the *de novo* assembly of complete genomes and some of the technologies have specific error characteristics (DiGuistini *et al*., 2009; Pop, 2009; Turner *et al*., 2009), hence Sanger sequencing is still superior in terms of data quality (Ansorge, 2009). For example, the 454 technology offer much longer reads than the Illumina or SOLiD methods, but its inability to accurately determine homopolymers longer than 3-4 bases remains a concern (Voelkerding *et al*., 2009). With Sanger sequencing it is also possible to generate read pairs that link distant regions of large genomes, by cloning large inserts and taking reads from both ends, which is not possible with some of the new technologies. Repeat sequences also pose a difficulty during data assembly if the read lengths are shorter than the repeat length, since the assembly algorithms are unable to determine the length of the repeat region. Because the read quality and error distribution for the new technologies are very different from Sanger methods new software tools are needed for processing and assembly. Progress has already been made in the development of algorithms for

the *de novo* assembly of very short reads (Whiteford *et al.*, 2005; Warren *et al.*, 2006; Sundquist *et al.*, 2007; Dohm *et al.*, 2007; DiGuistini *et al.*, 2009; Pop, 2009, Turner *et al.*, 2009), and software has also appeared which can incorporate Sanger sequencing data into next-generation sequence assemblies to improve the overall consensus quality (Goldberg *et al.*, 2006; Wicker *et al.*, 2006, Pop, 2009).

Despite all the obstacles, short single reads are still very useful for re-sequencing, because the reference sequences provide an essential backbone against which the short reads can be aligned uniquely. Approaches combining next-generation sequencing technologies with Sanger sequencing have proved to be successful in overcoming the systemic errors of a particular method and reducing costs. For example, DiGuistini and colleagues (2009) used Illumina, 454 and Sanger sequence data for the *de novo* assembly of a fungus genome sequence. The rapid improvement of the chemistries and algorithms has enabled the use of next-generation sequencing technology platforms for the genome sequencing and genome wide profiling of novel genetic variations in many different organisms, including bacteria (Holt *et al.*, 2008; Manning *et al.*, 2008; Qi *et al.*, 2009), plants (Bekal *et al.*, 2008; Novaes *et al.*, 2008), worms (Hillier *et al.*, 2008; Xia *et al.*, 2009) and humans (Wang *et al.*, 2008; Wheeler *et al.*, 2008). The next-generation technologies have reduced both the cost-per-reaction as well as the time required by orders of magnitude.

1.2.2. Identification of novel vaccine candidate genes from whole genome sequence data

The conventional approach for finding vaccine candidate genes has been to immunise animals with live infectious organisms, followed by the identification and purification of the immuno-reactive proteins and the determination of their amino acid sequences. The corresponding genes can then be identified and cloned for recombinant expression. The work is time-consuming and allows for the identification only of those antigens that can be purified in quantities suitable for

vaccine testing. Since the most abundant proteins are most often not good vaccine candidates it may take decades to develop a vaccine using the conventional approach.

1.2.2.1. Reverse vaccinology

The availability of complete genome sequences facilitates the design of vaccines starting from the *in silico* prediction of all antigens, independently of their abundance and without the need to grow the microorganism *in vitro*. The screening process usually involves the search for gene products with sequence or structural similarity to documented protective proteins or known microbial virulence factors (Ariel *et al*., 2003). A virulence factor is defined as any molecule produced by a pathogen that is essential for causing disease in a host (Finlay & Falkow, 1997). The genome-based selection of vaccine candidates, known as reverse vaccinology (Rappuoli, 2000), in combination with functional genomic studies, has been applied to several human pathogens, as illustrated by the following examples.

1.2.2.1.1. Neisseria meningitidis *vaccine candidates*

The first pathogen to which reverse vaccinology was applied was the Gram-negative bacterium, *Neisseria meningitidis*, a major cause of meningitis and bacterial septicaemia in children and young adults. From the 2,158 putative ORFs of the *N. meningitidis* type B MC58 genome (Tettelin *et al*., 2000), 570 sequences encoding potential surface-exposed or exported proteins were identified (Pizza *et al*., 2000). Of these, 350 proteins were successfully expressed, purified, and used to immunise mice. Using fluorescent-activated cell sorting (FACS) analysis, 91 proteins were shown to be surface-exposed and 28 were able to induce bactericidal antibodies. Seven surface-exposed antigens, which are conserved among the most prevalent *N. meningitidis* serogroups, are being evaluated as vaccine candidates (Grandi, 2003).

1.2.2.1.2. Streptococcus pneumoniae *vaccine candidates*

Current vaccines against *Streptococcus pneumoniae*, which causes bacterial sepsis, pneumonia and meningitis, have several limitations and are poorly efficacious in infants and the elderly. The

genome sequence of a serotype 4 strain of pneumococci was determined and 2,687 potential ORFs were identified (Wizemann *et al*., 2001). One hundred and thirty genes were selected, based on their predicted localization on the surface of the bacterium, and cloned for expression. Proteins predicted to be larger than 100 kDa were cloned in small subfragments to facilitate expression. The products of 108 ORFs or ORF fragments, comprising 97 genes, were expressed successfully and used to immunise mice. Six proteins conferred protection against disseminated *S. pneumoniae* infection and were shown to be both conserved within the species and immunogenic in humans. FACS analysis confirmed the surface localization of several of these antigens.

1.2.2.1.3. Chlamydia pneumoniae *vaccine candidates*

Chlamydia pneumoniae is an obligate intracellular bacterium and a common human pathogen with a biphasic life cycle similar to that of *E. ruminantium.* The cycle involves two developmental forms, spore-like infectious forms called elementary bodies and intracellular replicative forms known as reticulate bodies (Hatch, 1999). Montigiani and co-workers (2002) adopted a combined genomic-proteomic approach and identified 141 putative surface-associated proteins from the *C. pneumoniae* CWL029 genome (Kalman *et al*., 1999) by means of *in silico* analysis. Fifty-three of the selected proteins were confirmed to be surface-exposed by FACS analysis. Of these, 41 recognised a protein with the expected size on Western blots and 28 of the 53 antigens were identified on two-dimensional electrophoresis maps of elementary body extracts. Since a vaccine against *C. pneumoniae* requires, at least in part, to stimulate immune responses against proteins exposed on the surface of infectious chlamydiae, these data provide a way to a rational selection of new vaccine candidates.

1.2.2.1.4. Porphyromonas gingivalis *vaccine candidates*

Ross *et al*. (2001) aimed to identify previously unknown outer membrane proteins from the genome sequence of *Porphyromonas gingivalis*, a key periodontal pathogen, using a combination of global similarity searching, motif searching and intrinsic outer membrane probability. From

approximately 15,000 possible ORFs, 120 candidates were selected for the laboratory screen. One hundred and seven proteins were expressed successfully in *E. coli* and screened with antisera. Forty Western blot-positive proteins were purified and used to immunise mice that were subsequently challenged with live bacteria. Two antigens, both containing the OmpA motif, demonstrated significant protection.

1.2.2.1.5. **Bacillus anthracis** *vaccine candidates*

Bacillus anthracis, the causative agent of anthrax, is considered to be one of the most likely biological warfare agents. Using a reductive *in silico* selection strategy, Ariel and colleagues (2003) identified 520 candidates from a total of 5,054 predicted ORFs. They excluded ribosomal proteins, phage proteins, fragmented genes, and genes with more than two paralogs. ORFs with more than four predicted trans-membrane segments were also eliminated to avoid possible cloning problems. Surface-associated or secreted proteins and virulence-associated proteins were selected, namely: toxins, S-layer homology domain proteins, repeat proteins (tetratricopeptide, leucine-rich, Ankyrin, Collagen-like), adhesins or colonization factors, lytic enzymes and zinc proteases. Proteomic analysis of a *B. anthracis* membrane-associated fraction by two-dimensional gel electrophoresis was employed to demonstrate the expression and cellular location of the *in silico* selected chromosomal gene products and to identify immunogenic membrane or outer surface proteins. Close to 100 spots from the gel were analysed by matrix-assisted laser desorption ionization–time-of-flight mass spectrometry and were found to represent 32 proteins. Thirty-eight spots cross-reacted with sera from *B. anthracis* infected animals. Further analysis established that the cross-reactive spots, which represented the products of eight ORFs, were indeed expressed *in vivo* during exposure to *B. anthracis* and were able to elicit an immune response.

1.2.2.2. Comparative genomics

The complete genomic sequence from two or more isolates of the same species, or closely related species, allows for a detailed direct genome-to-genome comparison. In particular, the analysis of the genetic variability between pathogenic and closely related non-pathogenic microorganisms leads to the identification of genes potentially responsible for the acquisition of virulence. For example, Maione and colleagues (2005) analysed the genome sequences of eight Group B *Streptococcus* isolates and identified four proteins that proved to be highly protective against a large panel of strains.

1.2.2.3. Expression profiling

The advent of whole genome sequencing has also stimulated the development and widespread application of DNA microarray technology to study global changes in the expression of bacterial genes that are essential for pathogenesis and survival in the host. For example, DNA microarrays carrying the entire gene repertoire of *N. meningitidis* were used to identify protective antigens from genes that were regulated during interaction with human epithelial cells (Grifantini *et al*., 2002). In another study, virulence genes were identified by a DNA microarray-based comparison of a pathogenic and a nonpathogenic strain of *Xylella fastidiosa* (Koide *et al*., 2004). Microarray analysis was also successfully used to identify group A *Streptococcus* genes expressed during phagocytic interaction with human polymorphonuclear leukocytes (Voyich *et al*., 2003), and Merrell and colleagues (2002) selected new targets for cholera vaccine development by identifying the components that are required for the hyperinfectious state and dissemination of *Vibrio cholerae*.

Microarray data have limitations because mRNA levels do not necessarily correlate with protein expression levels (Debouck & Goodfellow, 1999). Expression of a transcribed gene may be regulated at the level of translation and protein products may be subject to control by posttranslational modifications. Currently there are also practical constraints to the use of this

technology to study intracellular bacteria, from which only small amounts of mRNA can be isolated.

New-generation sequencing of transcriptomes allows one to map and quantify transcripts in biological samples, an approach termed RNA-Seq (Nagalakshmi *et al*., 2008). Recent studies have shown that RNA-Seq is more accurate in quantifying transcripts than microarrays (Fu *et al*., 2009). Microarray data are restricted by the dynamic detection range of the scanner; background, saturation, and spot density and quality all influence the accuracy of the microarray data. In contrast, sequence data have a linear dynamic range only limited by the sequencing depth and allow for the detection of even extremely minimally expressed transcripts (Marguerat & Bähler, 2009; Tang *et al*., 2009; Van Vliet, 2010). For instance, Ozsolak and co-workers (2009) were able to sequence femtomole quantities of poly(A) *Saccharomyces cerevisiae* RNA. Although RNA-Seq has been successfully applied to studies of the transcriptomes of numerous eukaryotic genomes, only a few examples of bacterial transcriptome analysis have been reported. For instance, Passalacqua and colleagues (2009) used SOLiD and Illumina sequencing data for a comprehensive transcriptome analysis of *Bacillus anthracis*, and Perkins and co-workers (2009) utilised strand-specific cDNA sequencing with the Illumina Genome Analyzer to analyse the transcriptome of *Salmonella enterica* serovar Typhi.

1.3. AIMS OF THIS STUDY

The ultimate purpose of this study is to identify ORFs of *E. ruminantium* that induce strong cell-mediated immune responses for inclusion in a recombinant heartwater vaccine.

The first stage of the work was the completion and annotation of the entire genome sequence of the Welgevonden strain of *E. ruminantium* as presented in Chapter 2. Subsequently the metabolic pathways were analysed and compared to other organisms in the order Rickettsiales (Chapter 3) and the presence of an unusually large number of tandemly repeated and duplicated sequences

was investigated (Chapter 4). Chapter 5 describes the identification of potential vaccine candidates from the genome sequence using bioinformatic tools, the selection from among these of ORFs whose products induce cellular immune responses *in vitro*, and the evaluation of the vaccine candidates for their ability to stimulate protection against *E. ruminantium* infection in animal trials. The final chapter summarises the progress made during the course of this study, and makes suggestions for further investigation.

CHAPTER 2

The completion and annotation of the genome sequence of *Ehrlichia ruminantium* (Welgevonden)

2.1. INTRODUCTION

Two different strategies for generating whole genome sequences are frequently used (Frangeul *et al.*, 1999). The first strategy is the ordered-clone approach that uses large insert libraries to establish a contiguous set of overlapping clones covering the entire genome. Small insert libraries of the clones are then sequenced to obtain the complete genome sequence. The second strategy, direct shotgun sequencing (Bankier *et al.*, 1987), does not require preliminary data such as a physical map before starting the sequencing phase, and has therefore become the method of choice for sequencing small genomes.

In principle, a genome of arbitrary size may be directly sequenced by the shotgun method, provided that it can be uniformly sampled at random and that it does not contain long repeats. Shotgun sequencing has been successfully applied to the sequencing of larger and larger clones; from plasmids to cosmid clones (40 kb) (Edwards *et al.*, 1990), to artificial chromosomes cloned in bacteria and yeast (50-100 kb) (Wooster *et al.*, 1995), and bacterial genomes (1-2 Mb) (Fleischman *et al.*, 1995). Typically the strategy involves the construction of two DNA libraries: one library with relatively short inserts and a second library containing large inserts. The large fragments (20-300 kb) are usually cloned in phage lambda (λ), cosmids, or bacterial artificial chromosomes (BACs). The small insert library, with 1-2 kb fragments cloned into a plasmid or bacteriophage vector, is used for the bulk of the DNA sequencing and this is supplemented by sequences obtained from the larger fragments. Finally, the contigs are ordered and the remaining sequence gaps are closed by primer walking, primarily from linking clones in the second library (Frangeul *et al.*, 1999).

Once a DNA sequence has been completed, the annotation phase begins. The aim of annotation is to identify primary structural features within the DNA sequence, including the identification of ORFs and the analysis of possible terminator structures and promotors. A typical bacterial translational start site consists of a Shine-Dalgarno sequence, or ribosomal binding site (RBS), followed within 4-10 base pairs by one of the start codons (ATG, TTG or GTG). An ORF ends with any of the three stop codons TAA, TGA, or TAG (Weaver & Hedrick, 1992). Functional predictions can be made by performing homology analysis. When an amino acid sequence displays a high level of similarity to a sequence with a known function from another organism, it is likely that the putative protein performs the same, or a similar, function. The identification of functional domains or motifs can also aid in determining the putative function of a gene. Examples of such domains include ATPase domains characteristic of ABC transporters (Higgins, 2001), helix-turn-helix domains which indicate DNA-binding (Brennan & Matthews, 1989), signal sequences typical of exported proteins (Von Heijne, 1985) and transmembrane helices (Von Heijne, 1992).

This study reports on the complete genome sequence of the Welgevonden isolate of *E. ruminantium* which was obtained from an *Amblyomma hebraeum* tick collected near the Onderstepoort Veterinary Institute, in Gauteng Province, South Africa (Du Plessis, 1985). This is the geographical area from which the original *Rickettsia ruminantium* was obtained (Cowdry, 1925a), which is one reason for designating this isolate as the type specimen of *E. ruminantium* (Dumler *et al.*, 2001). A physical map of the Welgevonden genome was constructed by De Villiers and others (2000). They estimated that *E. ruminantium* had a circular chromosome of approximately 1,576 kb in size, and nine previously published genes or cloned DNA fragments were located on the physical map. Two *E. ruminantium* libraries were constructed in lambda vectors; a λZAPII expression library, with an average insert size of 3 kb (Brayton *et al.*, 1997b) and a large insert library (15-23 kb) in LambdaGEM®-11 (Promega) (Van Heerden *et al.*, 2004a). An additional small insert library (600-1,500 bp) was constructed in a plasmid vector. These resources enabled us to complete the sequencing of the entire genome of *E. ruminantium* by

whole-genome shotgun sequencing.

After the completion of the *E. ruminantium* Welgevonden genome, but before the completion of the work described in this thesis, two other *E. ruminantium* genome sequences were published (Frutos *et al.*, 2006): one was from the Gardel strain which was isolated on the Caribbean island of Guadeloupe, this strain was designated Erga; the other was of a daughter strain of the original South African Welgevonden strain which had been subjected to 11–13 passages over 18 years in a different cell-culture environment, this strain was designated Erwe. These two sequences were not included in any of our analyses, partly because they only became available during the course of our work, and partly because Frutos and co-workers have published detailed comparisons of all three *E. ruminantium* sequences (Frutos *et al*., 2006, 2007).

2.2. MATERIALS AND METHODS

See Appendix B for materials and media components.

2.2.1. Genome sequencing and assembly

2.2.1.1. DNA extraction

Genomic DNA was prepared from the Welgevonden stock of *E. ruminantium* (Du Plessis, 1985) grown in a bovine aorta endothelial cell line as previously described (Van Heerden *et al.*, 2004b). Briefly, the DNA was extracted by purifying the elementary bodies on discontinuous Percoll density gradients (Mahan *et al.*, 1995). RNA and eukaryotic DNA were removed by treating the elementary bodies with RNase (100 mg/ml) and DNase I (150 mg/ml) for 1.5 h at 37°C. The RNase and DNase I were inactivated by adding 12.5 mM EDTA. Elementary bodies were washed with sterile water and lysed for 2 h at 55°C in 0.1 M EDTA, 0.15 M NaCl, 1.5% SDS and 300 µg/ml proteinase K. Genomic DNA was extracted from the resulting lysate using the phenol/chloroform/isoamyl alcohol (25:24:1) extraction method (Sambrook *et al.*, 1989).

2.2.1.2. Construction of small insert libraries

The bulk of the genome sequence was obtained by shotgun sequencing of clones from two small insert *E. ruminantium* (Welgevonden) genomic libraries. Initial sequencing was performed using an existing expression library, designated WL1, constructed in λZAPII by the ligation of *E. ruminantium* genomic DNA partially digested with *Sau*3A (Brayton *et al.*, 1997b). A second small insert library, designated WL3, was constructed in a plasmid vector as follows. Genomic DNA (30 µg) was nebulised in a Medel jet nebuliser reservoir (Medel, Italy) for 2 min at 100 kPa and fragments in the 600-1,500 bp range were selected by agarose gel electrophoresis. The ends of the fragments were filled in with Klenow Fill-In Kit (Stratagene) and subcloned into pMOS*Blue* (Amersham Biosciences) as specified by the suppliers. Ligation reaction products were precipitated and transformed into high efficiency XL1-Blue electroporation competent cells. The library was plated onto BioAssay plates at approximately 1,000 cfu per plate and colonies were lifted onto nitrocellulose membranes soaked in LB/glycerol and stored at -80°C.

2.2.1.3. Template preparation for DNA sequencing

Cloned inserts from the WL1 library were amplified with the standard T7 primer (5' GTA ATA CGA CTC ACT ATA GGG C 3') and primer WL1F (5' GCT CTA GAA CTA GTG GAT CCC 3'). PCR reactions were performed in 50 µl volumes, containing 5 µl of the phage supernatant, PCR buffer, 2.5 mM $MgCl_2$, 0.2 mM of each dNTP, 0.25 µM of each primer and 1.25 U Amplitaq Gold polymerase (Applied Biosystems). The temperature profile of the reactions, performed on a GeneAmp PCR System 9700 (Perkin-Elmer Applied Biosystems), was: initial denaturation of 20 min at 94°C; 10 cycles of 20 s at 94°C, 30 s at 58°C and 1 min 30 s at 72°C; 6 cycles of 20 s at 94°C, 30 s at 58°C and 2 min 30 s at 72°C; 6 cycles of 20 s at 94°C, 30 s at 58°C and 5 min at 72°C; 6 cycles of 20 s at 94°C, 30 s at 58°C and 7 min 30 s at 72°C; and a final extension for 10 min at 68°C. The PCR products were analysed on 1% agarose gels. Amplicons larger than 500 bp were selected and purified with either the Concert Rapid PCR Purification System (Gibco BRL Products) or the QIAquick PCR Purification Kit (Qiagen) using the protocols provided by the manufacturers.

Plasmid DNA from the WL3 clones was extracted using the QIAprep 8 Miniprep Kit and QIAvac 6S vacuum manifold (Qiagen) according to the manufacturer's instructions. Alternatively, the plasmid DNA was amplified with the TempliPhi™ DNA Sequencing Template Amplification Kit (Amersham Biosciences). Plasmid DNA was subsequently digested with *Eco*RI (Roche) and *Xba*I (Roche) to determine insert sizes, and plasmids with inserts larger than 400 bp were sequenced.

2.2.1.4. DNA sequencing

All sequencing was performed using the Dye-terminator Cycle Sequencing kit (Applied Biosystems) on an ABI Prism 377 DNA sequencer or an ABI3100 Genetic Analyzer (Perkin Elmer Applied Biosystems) according to the protocols recommended by the manufacturer. WL1 clones were sequenced with either the WL1F primer (inserts 500-700 bp) or both the WL1F and

standard T7 primers (inserts > 700 bp), whereas WL3 clones with insert sizes smaller than 700 bp were sequenced with the pMOS_T7 primer (5' TAA TAC GAC TCA CTA TAG GG 3') and sequences of those with inserts larger than 700 bp were determined from both ends with pMOS_T7 as well as the universal M13(-47)F primer (5' CGC CAG GGT TTT CCC AGT CAC GA 3').

2.2.1.5. Sequencing data analysis and assembly

The Staden sequence analysis package (Staden *et al.*, 2000) was used for sequence analysis and assembly. Sequences were base called with Phred (Ewing *et al.*, 1998) and assembled using PHRAP (P. Green, http://www.phrap.org/). The following PHRAP parameters were used, forcelevel 1, minimum alignment score of 50, and minimum length of matching word equal to 20. GAP4 (Bonfield *et al.*, 1995) was used for manual checking and editing. Existing sequences of selected large clones (Louw *et al.*, 2002; Pretorius *et al.*, 2002a; Van Heerden *et al.*, 2002) from the LambdaGEM®-11 library were also added to the assembly.

2.2.1.6. Gap closure and quality assessment

Initial contig ordering was performed by exploiting synteny with the preliminary genomic sequence of *Ehrlichia chaffeensis*, the closest relative of *E. ruminantium* for which genomic data were available at that time. The preliminary *E. chaffeensis* sequence was made available by The Institute for Genomic Research (www.tigr.org). The remaining gaps were filled by performing PCR amplification using all combinations of primers designed to anneal to the ends of all contigs. All primers were designed with annealing temperatures of 50-55°C. The PCR reactions contained 25 ng genomic DNA, PCR buffer, 0.25 µM of each primer, 0.2 mM dNTPs and 1 U TaKaRa Ex Taq™ (TaKaRa Bio Inc.). Amplification was carried out under the following conditions: one cycle at 94°C (5 min), followed by 30 cycles at 94°C (10 s), 50°C (30 s) and 72°C (30 s), and a final extension at 72°C (7 min). When more than one amplicon was obtained, the PCR was repeated at a higher annealing temperature (53.5°C). PCR reactions which produced no

amplification product were repeated at an annealing temperature of 48°C. Repeat regions, areas represented by single reads or clones, and regions of low quality were resequenced from PCR products generated from *E. ruminantum* (Welgevonden) genomic DNA. In total we designed and used 852 primers for gap closure (Appendix C1).

Particular attention was paid to ensuring the accuracy of the final sequence and all contigs were carefully examined to identify problems in the sequence. These problems included gaps in the sequence, weakly supported sequence, ambiguities in the sequence, and sequence on only one strand. The minimal criteria were established as either to obtain unambiguous sequence on both strands or, if sequence was available on only one strand, this had to be unambiguously confirmed on multiple clones, preferably from more than one library. The electropherogram data were used to edit sequences visually and, where discrepancies could not be resolved or a clear assignment made, the templates were resequenced or PCR amplicons were generated to obtain data of high quality. The same procedures were followed to check potential frameshifts, apparent chimeric sequences and areas containing repeats.

The integrity of the assembly was validated by comparing the positions of mapped genes and restriction sites to the physical map of De Villiers *et al*. (2000). A computed restriction map was created using the Staden package program Spin (Staden *et al*., 2000) and the recognition sites of the endonucleases *Ksp*I, *Rsr*II and *Sma*I.

2.2.2. Annotation and analysis

2.2.2.1. Selection of a gene set

The potential protein-coding genes were assigned by a combination of computer prediction and similarity searching. Three gene modelling programs, GeneMarkS (Besemer *et al.*, 2001), Orpheus (Frishman *et al.*, 1998) and Glimmer (Delcher *et al.*, 1999), were used independently to predict potential protein coding sequences (CDSs). RBSfinder (http://www.tigr.org/software/)

was used to assist with the location of start codons. When more than one potential start codon was identified, the first was arbitrarily chosen for annotation. The GC content, correlation scores and codon usage graphs from the Artemis sequence viewer and annotation tool (Rutherford *et al.*, 2000) were also taken into consideration to select a gene set. Each CDS in the gene set was given a systematic identification number, starting with Erum0010.

In parallel, the entire genome sequence was used to search non-redundant protein databases (GenBank and Swiss-Prot/TrEMBL) with the BLASTx program (Altschul *et al.*, 1997) to identify genes which were missed by the prediction algorithms. Transfer RNAs (tRNAs) were identified by tRNAscan-SE (Lowe & Eddy, 1997). If potential ORFs were partially or entirely overlapping, those showing similarity with known genes were chosen, and the longest one was selected unless the function of the shorter one was well supported in the databases.

2.2.2.2. Similarity searches and domain identification

Proteins predicted from the revised gene set were searched against non-redundant protein databases using FASTA (Pearson, 2000) and BLASTp (Altschul *et al.*, 1997). Domain analysis of predicted proteins was performed by searching Pfam (Bateman *et al.*, 2004) and PROSITE (Sigrist *et al.*, 2002). Mreps (Kolpakov *et al.*, 2003) and Tandem Repeats Finder (Benson, 1999) were used to detect tandem repeats. The results of all searches were assembled and predicted proteins were manually annotated in Artemis. Addresses of web based programs used in this study can be found in Appendix F.

Regions of the genome were assigned for analysis to eight different annotators, each of whom adhered to a set of rules created in order to keep the annotation as consistent as possible. First, each identified region was assigned a gene name, gene product, class and colour. Gene names followed the Demerec standard (Demerec *et al.*, 1966), consisting of a unique three-letter abbreviation intended to imply a function, followed by a capital letter to distinguish different genes related to the same function. The names of duplicated genes were followed by a number which indicated their order in the genome. We consulted Gene Ontology terminology (The Gene

Ontology Consortium, 2000) for the definition of gene products, and for functional classification we used the protein classification scheme created for *E. coli* (Riley, 1993) (Appendix D). For proteins where there was not enough evidence to be certain of the functional designation we used either "probable", for those that we believed were likely to be correct, or "possible" for those in which we were less confident. Predicted proteins with unknown functions were placed into one of two categories: "unknown" was used for ORFs that had no informative data (including similarity to genes of known function, matches to Pfam or PROSITE entries, or informative hydrophobicity plots), and "conserved hypothetical protein" was used for ORFs that had matches to other proteins of unknown function. An Enzyme Commission (EC) number (http://www.chem.qmul.ac.uk/iubmb/enzyme/) was allocated to predicted proteins homologous with proteins having an identified enzymatic function. Fasta, Pfam and Prosite matches and other motifs (transmembrane helices, signal sequences and helix-turn-helix motifs) were also included. Additional descriptive information, for example repeat sequences and self matches, was added when it was deemed to be useful. Pseudogenes were defined as regions with stop codons that interrupted reading frames, these were typically detected among the BLASTx results. Finally the author and one other annotator reviewed and standardised the entire annotation. The complete annotated sequence data were submitted to the European Molecular Biology Laboratory data bank under accession no. CR767821. A more detailed version of the annotation, together with supplementary information, can be downloaded in Artemis-compatible format from http://www.bi.up.ac.za/Ehrlichia_ruminantium/.

2.2.2.3. Subcellular localisation prediction of ORFs

SignalP (Nielsen *et al.*, 1997), TMHMM2.0 (Krogh *et al.*, 2001) and Phobius (Käll *et al.*, 2004) were used to detect putative signal peptides and transmembrane helices. We used PSORTb2.0 (Gardy *et al.*, 2005) and CELLO (Yu *et al.*, 2004) to assign proteins to likely subcellular locations.

2.3. RESULTS AND DISCUSSION

2.3.1. Sequence determination of the entire genome

2.3.1.1. Library construction

A major technical difficulty was the inability to construct *E. ruminantium* libraries in vectors that have proved efficient for other bacterial DNA. Brayton and co-workers created an *E. ruminantium* large insert library in a cosmid vector (Brayton *et al*., 1999) and they found that the *E. ruminantium* clones were unstable in the SuperCos1 vector and most clones did not grow reproducibly. Clones containing AT-rich inserts have been found to be difficult to grow by other investigators (Reddy, 1995; Pan *et al*., 1999; Gardner *et al*., 2002). Brayton and colleagues speculated that the lower melting temperature of AT-rich clones decreases their stability during growth at 37^{o}C or that the clones are targeted as intruders by the host cells because of the difference in AT content between the clone and the host cells. It has also been shown that *E. ruminantium* promotors are active in *E. coli* (Van Vliet *et al*., 1994; Brayton *et al*., 1997b) and it is believed that the expression of certain *E. ruminantium* genes suppresses host cell growth. Difficulties with cosmid libraries have been reported by other workers: high expression levels of *Bacillus subtilis* genes were toxic to the *E. coli* host cells (Kunst *et al*., 1997); under-representation of certain regions of the chromosome and unstable inserts were found in *Mycobacterium tuberculosis* cosmid libraries (Brosch *et al*., 1998); and a cosmid library of the *Sulfolobus solfataricus* genome covered only 70% of the chromosome (She *et al*., 2000). These limitations can be overcome by using low-copy-number vectors, such as BACs, although the laborious construction of BAC libraries can be a drawback (Frangeul *et al*., 1999).

Our λZAPII library, prepared using a partial *Sau*3A digest of *E. ruminantium* genomic DNA, was also found to have limitations; it was not completely random and it contained chimeric clones. After sequencing ~3,000 clones we only had about one genome equivalent, all in small contigs, and it seemed unlikely that the genome sequence could be completed by sequencing more WL1 clones. A new library was therefore constructed in a plasmid vector and the DNA used to make

this library was nebulised instead of being cut with restriction enzymes, since it is believed that mechanical shearing maximises the randomness of the DNA fragments. We selected a narrow fragment size range, between 600 bp and 1,500 bp, to minimise variations in the growth of different clones. In addition, we chose the 1,500 bp limit to minimise the number of complete genes that might be present in a single fragment, in the hope that this would reduce the chance of clone losses as a result of the expression of deleterious gene products. The plasmid library, designated WL3, had an average insert size of 700 bp. Although it contained some chimeric clones it was more representative than the lambdaZAPII library and provided sufficient sequence data to complete the genome sequence.

2.3.1.2. Genome assembly

We used the random shotgun approach to assemble the entire genome sequence of *E. ruminantium*. A total of 21,206 random sequence reads were assembled to generate a draft sequence consisting of 511 contigs with an average length of 3,318 bp and a total contig length of 1.7 Mb. Only 97 of the contigs were larger than 5 kb, of which 60 contigs were 5 to 10 kb in length and 37 were more than 10 kb in length.

Finishing was carried out by visually editing the sequences in all contigs, followed by gap closing. We manually scanned through the assembled contigs and noted regions where we were dissatisfied with the supporting reads. A large proportion of these regions were composed of tandem repeats and dispersed repeat units, some up to several hundreds of base pairs in length. Such repeats are very difficult for assembly engines to handle, since no single sequencing read covers the entire repetitive element. Consequently all areas containing repeats were checked by PCR amplification of the complete repeat region and sequencing of the amplicons. We found that some dispersed repeats were incorrectly assembled and when these problems were resolved a number of gaps were closed. In a few instances we were unable to determine an absolute number

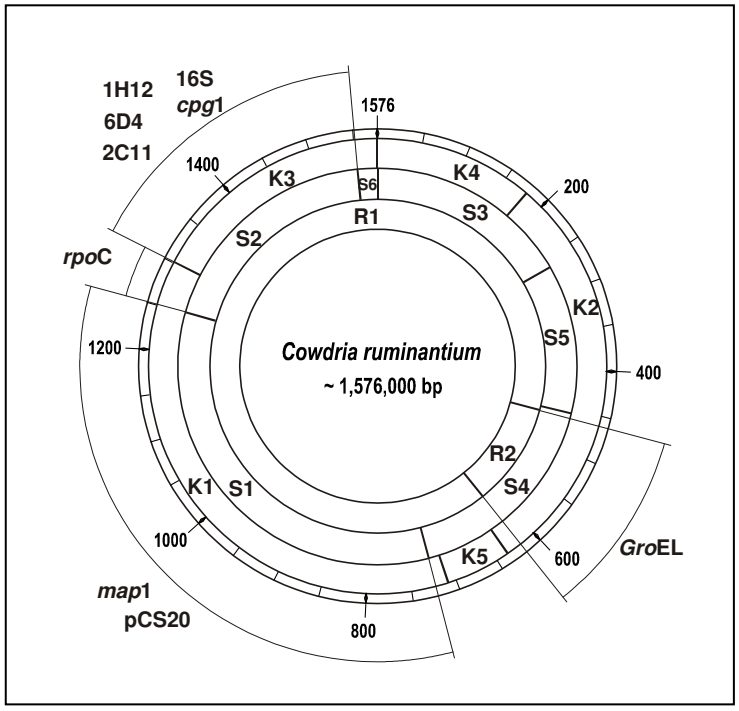
of tandemly repeated sequences; this was noted in the annotation. The repeat sequences will be discussed in Chapter 4.

As we were nearing completion of the finishing phase we found that we still had many small contigs that were not being incorporated into the assembly, almost all of which contained reads from the WL1 library. A BLAST search revealed that most of these sequences matched mycoplasmas and we concluded that this was the result of mycoplasma contamination in the cell cultures in which the *E. ruminantium* organisms were grown. The contamination of cell cultures with mycoplasmas is a very common problem (Langdon, 2004; Mariotti *et al*., 2008), which we too had experienced previously. We had, however, managed to eliminate this contamination before the construction of plasmid library WL3, hence mycoplasma clones were present only in the older WL1 library. In all, about 130 small contigs (368 reads) were omitted from the assembly.

During the finishing process we closed 143 gaps with an average gap size of 326 bp. In many instances (20%) there were in fact no physical gaps but the contig overlaps were too small to be recognised by the assembly algorithms as a reasonable join. Only 39 (28%) of the gaps were longer than 10 bp, the largest being 2,540 bp. The final assembly contained 25,648 reads with an average length of 569 bp, giving 9.6-fold coverage of the genome.

The final phase of the finishing process was global sequence validation. The structure of the assembled circular genome was confirmed by comparing a computer-generated restriction map based on the assembled sequence for the endonucleases *Ksp*I, *Rsr*II and *Sma*I, with the experimentally generated restriction map. The restriction fragments from the sequence-derived map matched those from the physical map in size and relative order (Figure 2.1). The positions of the mapped genes also correlate with their positions in the assembled genome.

A



B

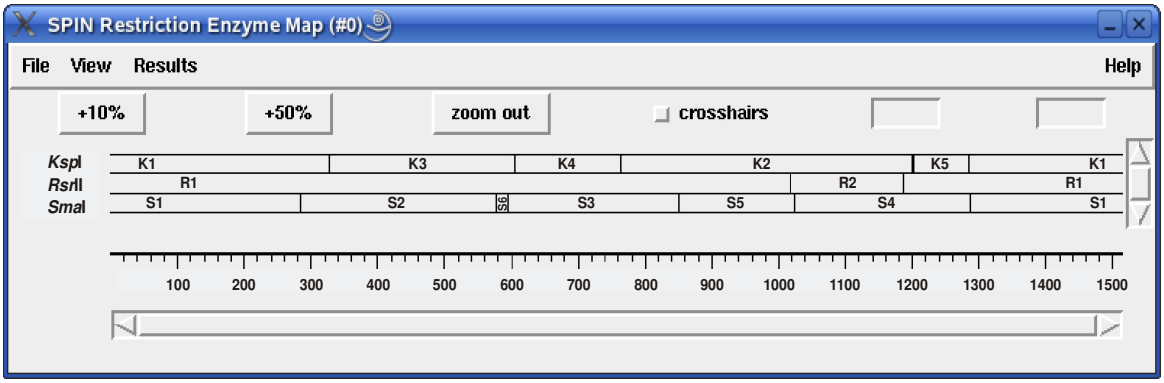


Figure 2.1. **A.** The physical map of De Villiers *et al*. (2000). From inside to outside, the circles represent the *Rsr*II, *Sma*I and *Ksp*I restriction sites respectively. The outer circle illustrates the scale in kilobases. Mapped genes and clones are also indicated. **B.** A computer-generated restriction map of the completed *E. ruminantium* genome sequence, showing the cutting sites of the endonucleases *Ksp*I, *Rsr*II and *Sma*I. The scale of the map is shown in kilobases. Base pair 1 on the physical map correlates with position 605,342 on the computer-generated map.

2.3.2. Annotation of the *E. ruminantium* genome sequence

2.3.2.1. Assignment of potential coding regions

We identified 920 coding sequences with an average length of 1,032 bp, of which 32 (3.5%) probably represent pseudogenes (Table 2.1, Figure 2.2). The low protein coding capacity of the genome (62%) is even more extreme than that for the related pathogen *Rickettsia prowazekii*, which is 76% coding (Andersson *et al.*, 1998). Eighty-eight percent of the ORFs have an ATG start codon, 8% TTG, and 4% GTG. A total of 36 tRNA genes were identified, and since they are dispersed around the genome they are likely to be transcribed as single units. One set of ribosomal RNA (rRNA) genes and two small RNA-encoding genes, tmRNA and RNase P subunit B (*rnpB*), were assigned to the genome.

Table 2.1. General features of the genome of the Welgevonden strain of *E. ruminantium.* (From Collins *et al.,* 2005)

Size	1,516,355 bp
G+C content	27.5%
% Protein coding regions (not including pseudogenes)	62.0%
Total number of CDSs	920
average length	1,032 bp
Probable pseudogenes	32 (3.5%)
average length	276 bp
Predicted protein coding sequences	888 (96.5%)
average length	1,059 bp
CDSs with functional information*	758 (82.8%)
conserved hypothetical genes	50 (5.5%)
genes with no functional information	80 (8.7%)
Stable RNAs	
number of ribosomal RNAs	3
number of transfer RNAs	36
number of other RNAs (tmRNA, rnpB)	2
Simple sequence repeats	1,590 bp (0.1%)
Tandem repeats	82,146 bp (5.4%)
Dispersed repeats (direct and inverted)	45,397 bp (3.0%)
TOTAL	129,133 bp (8.5%)

* Includes CDSs with database matches to genes of known function, matches to Pfam or PROSITE entries, or informative hydrophobicity plots.

2.3.2.2. Functional assignment of protein-encoding genes

Translated amino acid sequences of 920 potential protein-encoding genes in the genome were compared with sequences in non-redundant databases. We could assign informative data to 758 CDSs: 520 (56.5%) were allocated a specific function, 175 (19.0%) were predicted to encode membrane-associated or exported proteins, and 63 (6.8%) could not be classified but had some miscellaneous information. Fifty CDSs (5.4%) were similar to conserved hypothetical genes of unknown function, and eighty (8.7%) did not show any sequence similarity to known genes in other organisms nor was any other functional information identified. Many of these unknown genes will probably have functions related to species specialisation. The putative protein-coding genes whose function could be anticipated were grouped into categories according to their different biological roles (Table 2.2, Figure 2.2, and Figure 2.3). On the gene map (Figure 2.3) the location, length and direction of the ORFs are indicated, with colour codes corresponding to functional categories. See Appendix E for a complete gene list with annotation. Obviously the genes assigned in this study merely represent the coding potential of the genome for proteins and RNAs under the defined assumptions, and the real gene assignment will eventually have to be confirmed experimentally.

2.3.2.3. General features of the genome

The circular genome of the Welgevonden strain of *E. ruminantium* is 1,516,355 bp in length with a low G+C content (27.5%). The genomes of many other endosymbionts and intracellular pathogens have a high A+T content and it has been suggested that this has resulted from the loss of repair and recombination machinery, such as the SOS, base-excision and nucleotide-excision systems (*uvrABC*) (Akman *et al.*, 2002). This theory is supported by the fact that the mismatch-repair enzymes in *E. ruminantium* are limited to *mutS* and *mutL*, and there is only one subunit (A) of the ultraviolet-induced DNA damage repair system (*uvrABC*).

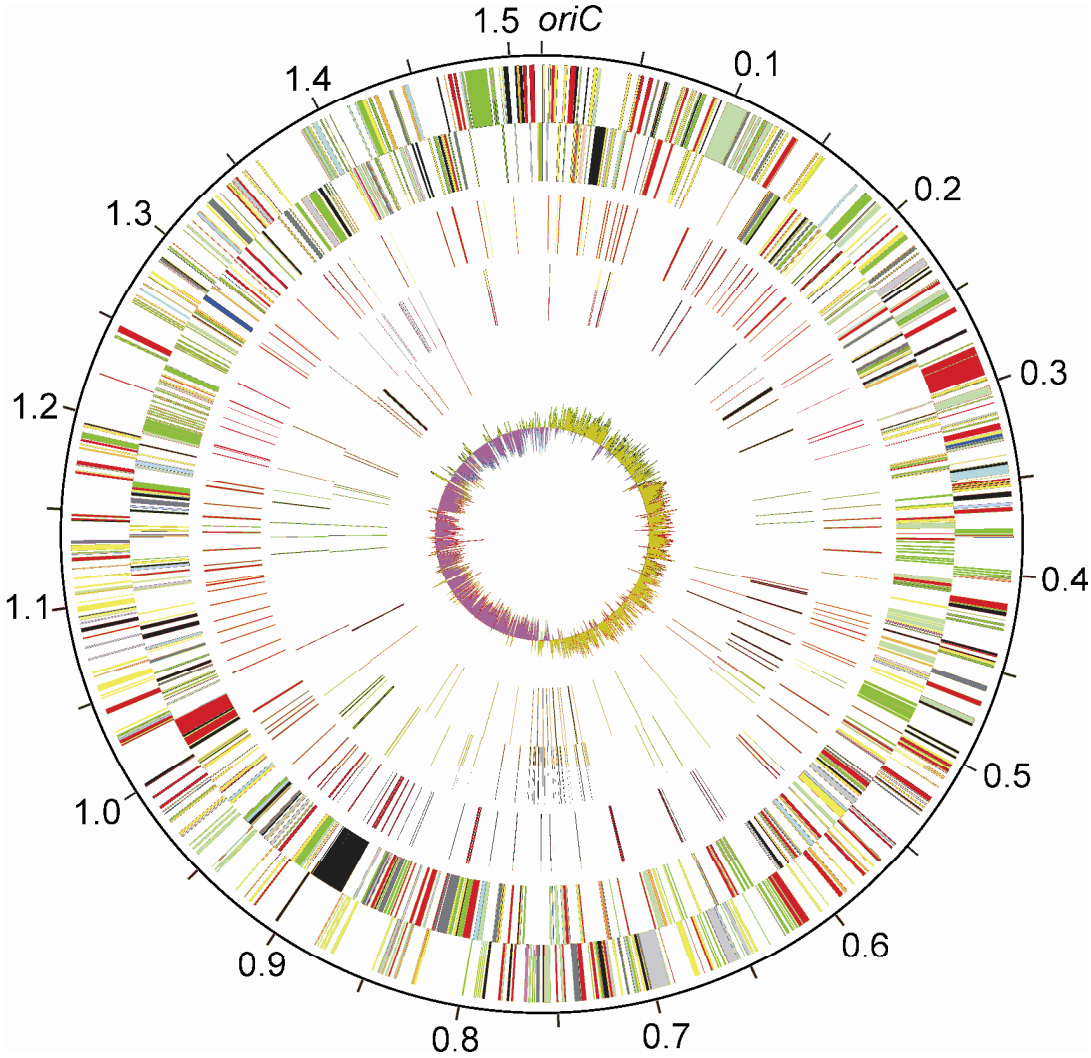
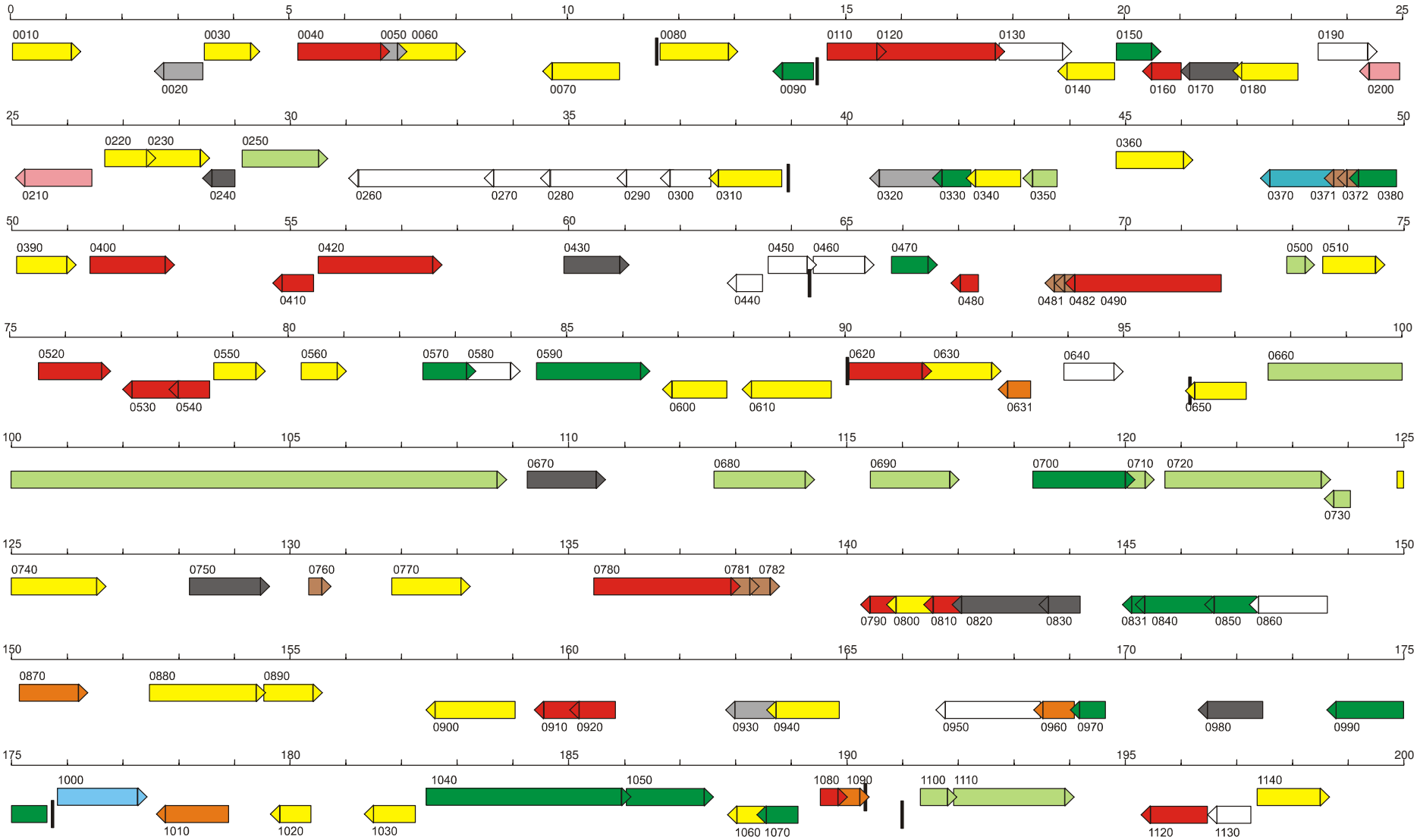
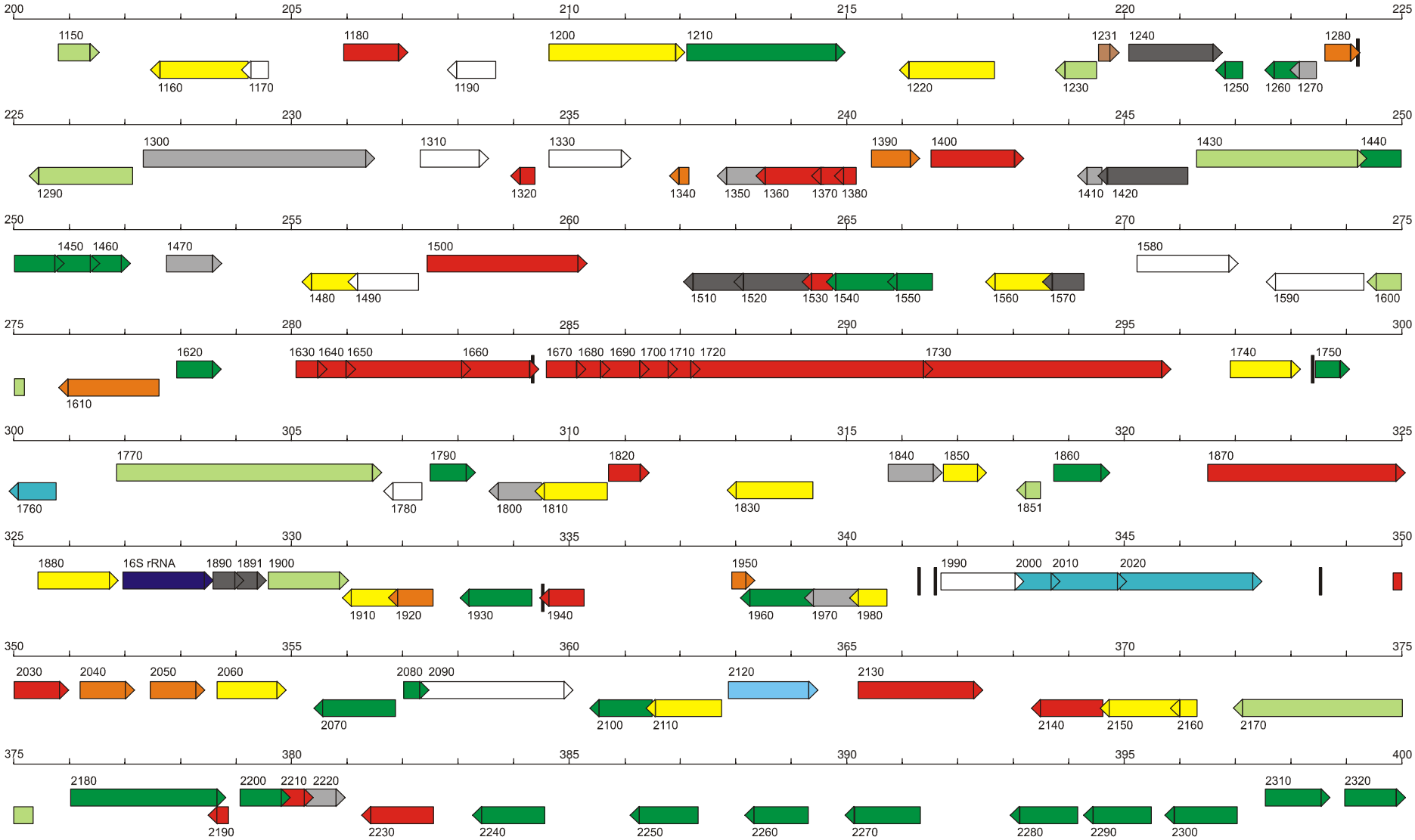


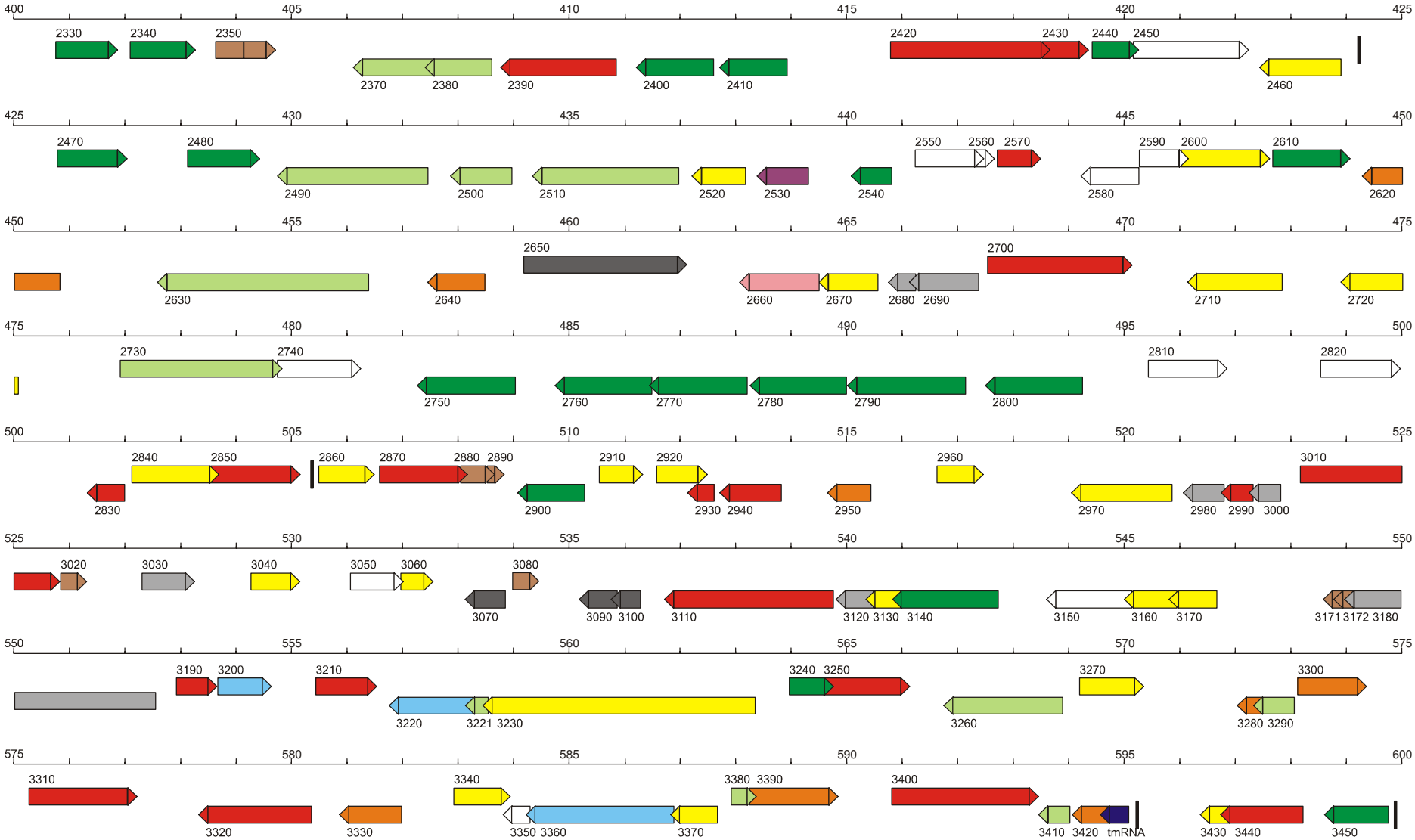
Figure 2.2. Circular representation of the genome of *E. ruminantium* (Welgevonden isolate). The outer circle indicates the scale in megabases. The remaining concentric circles are described from outside to inside. First and second circles, predicted coding sequences on the plus and minus strands respectively, colour-coded by function: dark blue, stable RNAs; black, chaperones and transporters; dark grey, energy metabolism; red, information transfer; yellow, central or intermediary metabolism; dark green, membrane and exported proteins; cyan, degradation of large molecules; purple, degradation of small molecules; pale blue, regulators; orange, conserved hypothetical proteins; pink, phage and insertion sequence elements; brown, pseudogenes; pale green, unknown; light grey, miscellaneous. Third circle, tandem repeats in red. Fourth and fifth circles, dispersed repeats (direct and inverted repeats) coloured in black. Sixth circle, G+C skew with values greater than zero in olive and less than zero in magenta. (From Collins *et al.,* 2005.)

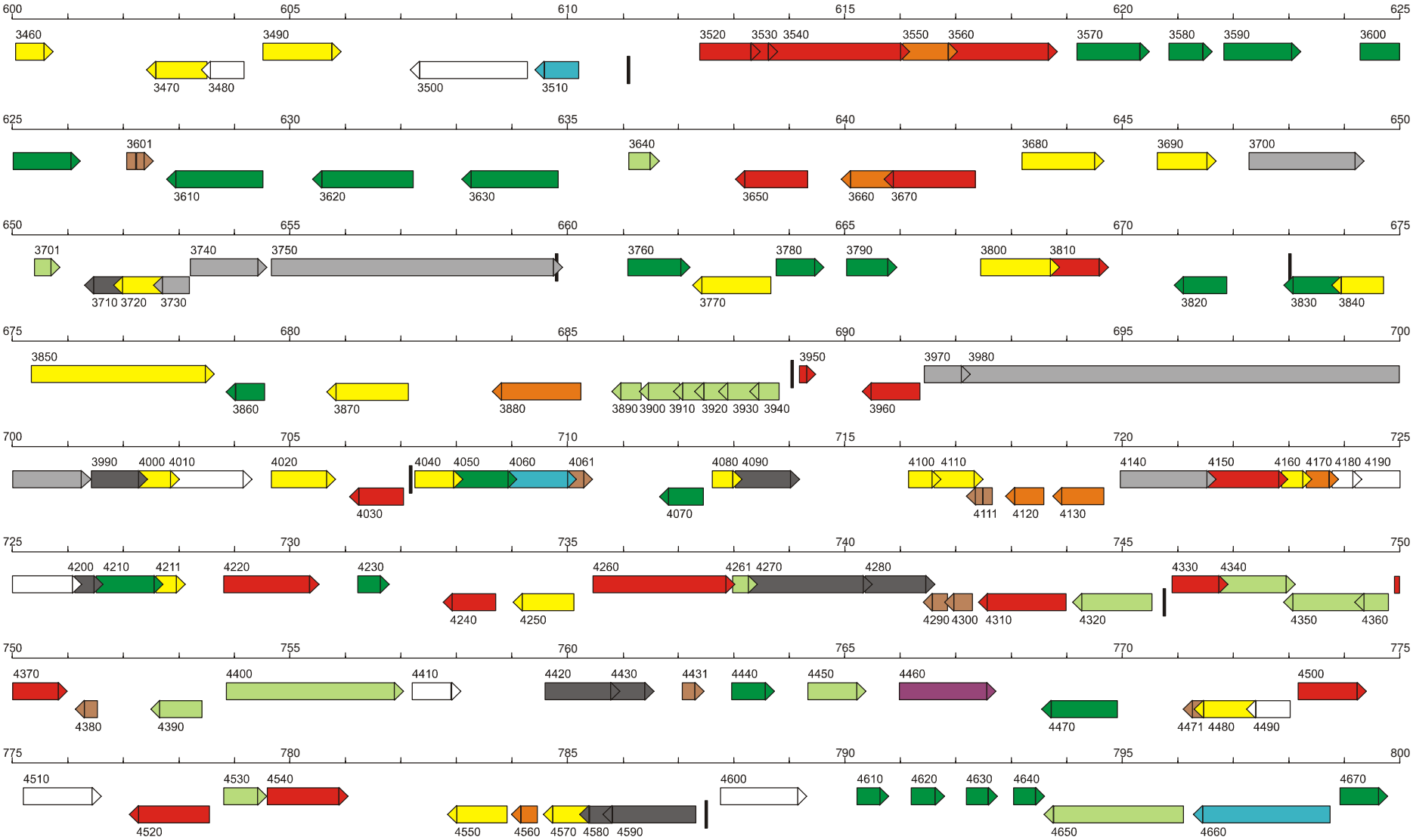
The origin of replication (*oriC*) has not been experimentally determined in *E. ruminantium*. In many other organisms there is a conserved arrangement of genes around *oriC* (Ogasawara & Yoshikawa, 1992), which is often located close to the *dnaA* gene, and a transition in GC-skew values is frequently evident at the origin and termination of replication (Lobry, 1996). In the *E. ruminantium* genome we found a clear shift in GC-skew values in two regions approximately 750 kb apart (Figure 2.2), but none of the genes normally associated with *oriC* were located near either of the transitions; in fact, except for *rmpH* and *rnpA*, such genes were not located near each other but were scattered throughout the genome. Comparisons of the closely related *Escherichia coli* K-12 and *Salmonella enterica* serovar Typhimurium genomes have revealed a high frequency of recombination in the terminus region which may be related to the mechanism of chromosome separation after replication (Hughes, 2000a). There are many duplications and translocations in the area around one of the shifts in GC-skew value (Figure 2.2), suggesting that this region has a higher rate of DNA reorganisation. This might indicate that the terminus of replication is located here, hence a position near the opposite transition in GC-skew values was chosen as base pair 1 of the genome. The *dnaA* gene was located at 506,593 bp, more than 200 kb away from the nearest transition in GC-skew values. Recently Ioannidis and co-workers (2007) suggested that the *oriC* region should be located 23 kb downstream, between Erum0180 and Erum0190, based on the presence of DnaA- and IHF-binding sites and the conservation of the boundary genes in related bacteria.

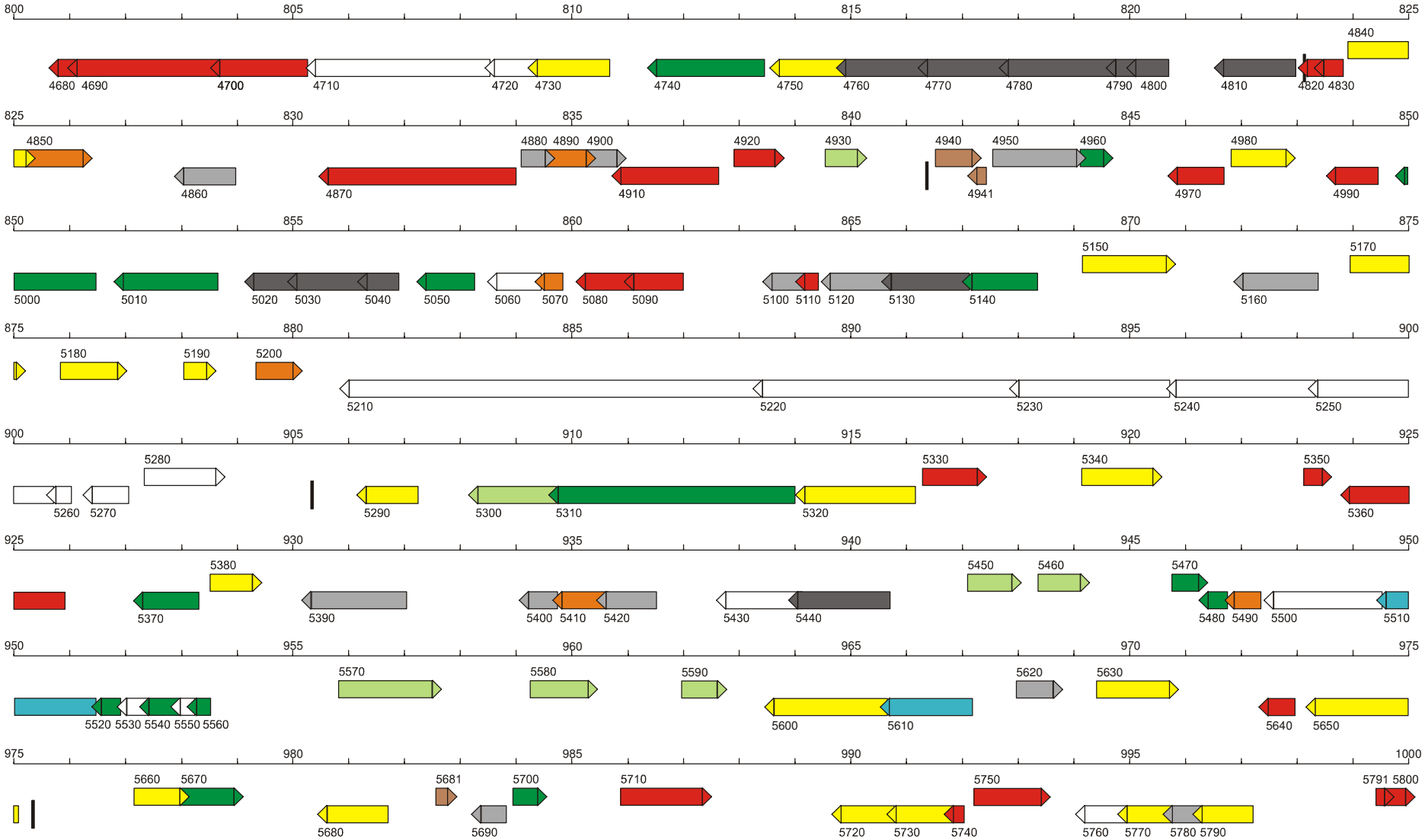
The unusual dispersion in *E. ruminantium* of genes normally found to be associated with *oriC* was also observed with other genes that normally occur in operons in other bacteria. One such example is the disruption of the ribosomal RNA (rRNA) operon: the 16S rRNA gene is located at 326,964 bp while the 5S and 23S rRNA genes are located on the opposite strand between 1,283,569 and 1,286,544 bp. Such unusual gene organisation patterns are a characteristic feature of intracellular bacteria (Andersson & Kurland, 1998) and are thought to be the result of recombination events that cause major chromosomal rearrangements which, in the isolated intracellular environment, cannot be corrected by recombination with other bacteria.

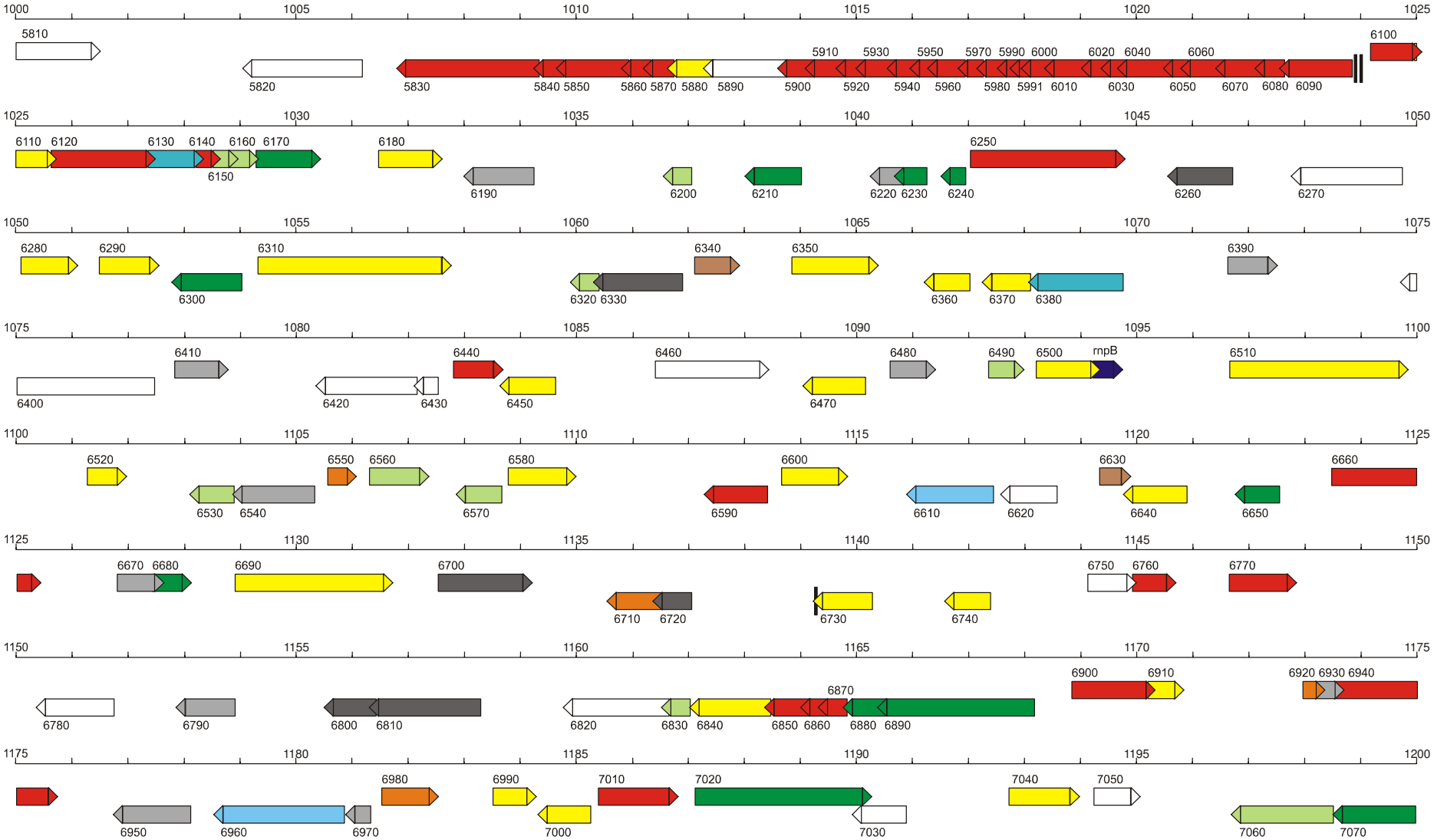

0 5 10 15 20 25
0010 0020 0030 0040 0050 0060 0070 0080 0090 0110 0120 0130 0140 0150 0160 0170 0180 0190 0200
25 30 35 40 45 50
0210 0220 0230 0240 0250 0260 0270 0280 0290 0300 0310 0320 0330 0340 0350 0360 0370 0371 0372 0380
50 55 60 65 70 75
0390 0400 0410 0420 0430 0440 0450 0460 0470 0480 0481 0482 0490 0500 0510
75 80 85 90 95 100
0520 0530 0540 0550 0560 0570 0580 0590 0600 0610 0620 0630 0631 0640 0650 0660
100 105 110 115 120 125
0670 0680 0690 0700 0710 0720 0730
125 130 135 140 145 150
0740 0750 0760 0770 0780 0781 0782 0790 0800 0810 0820 0830 0831 0840 0850 0860
150 155 160 165 170 175
0870 0880 0890 0900 0910 0920 0930 0940 0950 0960 0970 0980 0990
175 180 185 190 195 200
1000 1010 1020 1030 1040 1050 1060 1070 1080 1090 1100 1110 1120 1130 1140

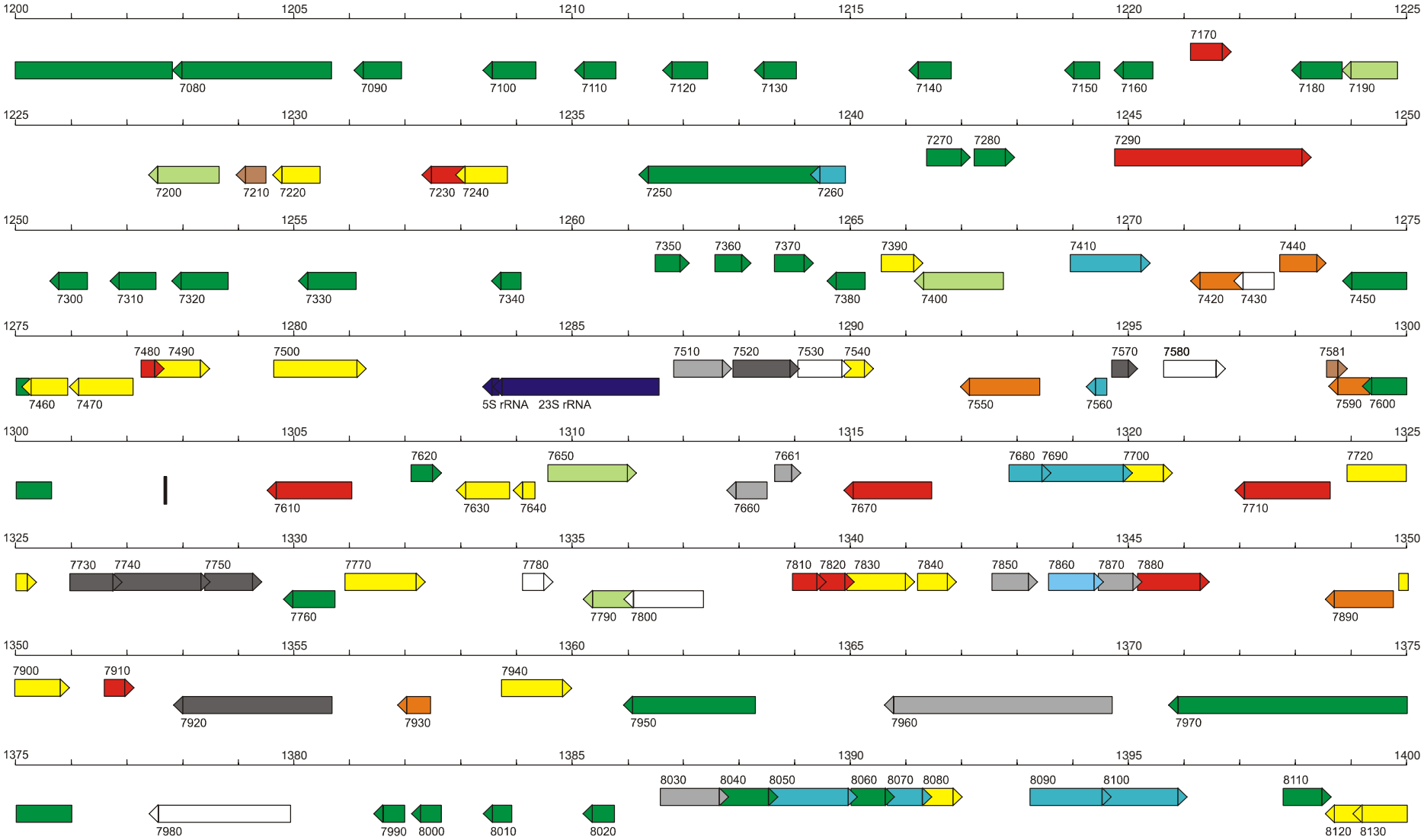

200 205 210 215 220 225
1150 1160 1170 1180 1190 1200 1210 1220 1230 1231 1240 1250 1260 1270 1280
225 230 235 240 245 250
1290 1300 1310 1320 1330 1340 1350 1360 1370 1380 1390 1400 1410 1420 1430 1440
250 255 260 265 270 275
1450 1460 1470 1480 1490 1500 1510 1520 1530 1540 1550 1560 1570 1580 1590 1600
275 280 285 290 295 300
1610 1620 1630 1640 1650 1660 1670 1680 1690 1700 1710 1720 1730 1740 1750
300 305 310 315 320 325
1760 1770 1780 1790 1800 1810 1820 1830 1840 1850 1851 1860 1870
325 330 335 340 345 350
1880 16S rRNA 1890 1891 1900 1910 1920 1930 1940 1950 1960 1970 1980 1990 2000 2010 2020
350 355 360 365 370 375
2030 2040 2050 2060 2070 2080 2090 2100 2110 2120 2130 2140 2150 2160 2170
375 380 385 390 395 400
2180 2190 2200 2210 2220 2230 2240 2250 2260 2270 2280 2290 2300 2310 2320


400 405 410 415 420 425
2330 2340 2350 2370 2380 2390 2400 2410 2420 2430 2440 2450 2460
425 430 435 440 445 450
2470 2480 2490 2500 2510 2520 2530 2540 2550 2560 2570 2580 2590 2600 2610 2620
450 455 460 465 470 475
2630 2640 2650 2660 2670 2680 2690 2700 2710 2720
475 480 485 490 495 500
2730 2740 2750 2760 2770 2780 2790 2800 2810 2820
500 505 510 515 520 525
2830 2840 2850 2860 2870 2880 2890 2900 2910 2920 2930 2940 2950 2960 2970 2980 2990 3000 3010
525 530 535 540 545 550
3020 3030 3040 3050 3060 3070 3080 3090 3100 3110 3120 3130 3140 3150 3160 3170 3171 3172 3180
550 555 560 565 570 575
3190 3200 3210 3220 3221 3230 3240 3250 3260 3270 3280 3290 3300
575 580 585 590 595 600
3310 3320 3330 3340 3350 3360 3370 3380 3390 3400 3410 3420 tmRNA 3430 3440 3450


600 605 610 615 620 625
3460 3470 3480 3490 3500 3510 3520 3530 3540 3550 3560 3570 3580 3590 3600
625 630 635 640 645 650
3601 3610 3620 3630 3640 3650 3660 3670 3680 3690 3700
650 655 660 665 670 675
3701 3710 3720 3730 3740 3750 3760 3770 3780 3790 3800 3810 3820 3830 3840
675 680 685 690 695 700
3850 3860 3870 3880 3890 3900 3910 3920 3930 3940 3950 3960 3970 3980
700 705 710 715 720 725
3990 4000 4010 4020 4030 4040 4050 4060 4061 4070 4080 4090 4100 4110 4111 4120 4130 4140 4150 4160 4170 4180 4190
725 730 735 740 745 750
4200 4210 4211 4220 4230 4240 4250 4260 4261 4270 4280 4290 4300 4310 4320 4330 4340 4350 4360
750 755 760 765 770 775
4370 4380 4390 4400 4410 4420 4430 4431 4440 4450 4460 4470 4471 4480 4490 4500
775 780 785 790 795 800
4510 4520 4530 4540 4550 4560 4570 4580 4590 4600 4610 4620 4630 4640 4650 4660 4670


800 805 810 815 820 825
4680 4690 4700 4710 4720 4730 4740 4750 4760 4770 4780 4790 4800 4810 4820 4830 4840
825 830 835 840 845 850
4850 4860 4870 4880 4890 4900 4910 4920 4930 4940 4941 4950 4960 4970 4980 4990
850 855 860 865 870 875
5000 5010 5020 5030 5040 5050 5060 5070 5080 5090 5100 5110 5120 5130 5140 5150 5160 5170
875 880 885 890 895 900
5180 5190 5200 5210 5220 5230 5240 5250
900 905 910 915 920 925
5260 5270 5280 5290 5300 5310 5320 5330 5340 5350 5360
925 930 935 940 945 950
5370 5380 5390 5400 5410 5420 5430 5440 5450 5460 5470 5480 5490 5500 5510
950 955 960 965 970 975
5520 5530 5540 5550 5560 5570 5580 5590 5600 5610 5620 5630 5640 5650
975 980 985 990 995 1000
5660 5670 5680 5681 5690 5700 5710 5720 5730 5740 5750 5760 5770 5780 5790 5791 5800


1000 1005 1010 1015 1020 1025
5810 5820 5830 5840 5850 5860 5870 5880 5890 5900 5910 5920 5930 5940 5950 5960 5970 5980 5990 5991 6000 6010 6020 6030 6040 6050 6060 6070 6080 6090 6100
1025 1030 1035 1040 1045 1050
6110 6120 6130 6140 6150 6160 6170 6180 6190 6200 6210 6220 6230 6240 6250 6260 6270
1050 1055 1060 1065 1070 1075
6280 6290 6300 6310 6320 6330 6340 6350 6360 6370 6380 6390
1075 1080 1085 1090 1095 1100
6400 6410 6420 6430 6440 6450 6460 6470 6480 6490 6500 rnpB 6510
1100 1105 1110 1115 1120 1125
6520 6530 6540 6550 6560 6570 6580 6590 6600 6610 6620 6630 6640 6650 6660
1125 1130 1135 1140 1145 1150
6670 6680 6690 6700 6710 6720 6730 6740 6750 6760 6770
1150 1155 1160 1165 1170 1175
6780 6790 6800 6810 6820 6830 6840 6850 6860 6870 6880 6890 6900 6910 6920 6930 6940
1175 1180 1185 1190 1195 1200
6950 6960 6970 6980 6990 7000 7010 7020 7030 7040 7050 7060 7070


1200
1205
1210
1215
1220
1225
7080
7090
7100
7110
7120
7130
7140
7150
7160
7170
7180
7190
1225
1230
1235
1240
1245
1250
7200
7210
7220
7230
7240
7250
7260
7270
7280
7290
1250
1255
1260
1265
1270
1275
7300
7310
7320
7330
7340
7350
7360
7370
7380
7390
7400
7410
7420
7430
7440
7450
1275
1280
1285
1290
1295
1300
7460
7470
7480
7490
7500
5S rRNA
23S rRNA
7510
7520
7530
7540
7550
7560
7570
7580
7581
7590
7600
1300
1305
1310
1315
1320
1325
7610
7620
7630
7640
7650
7660
7661
7670
7680
7690
7700
7710
7720
1325
1330
1335
1340
1345
1350
7730
7740
7750
7760
7770
7780
7790
7800
7810
7820
7830
7840
7850
7860
7870
7880
7890
1350
1355
1360
1365
1370
1375
7900
7910
7920
7930
7940
7950
7960
7970
1375
1380
1385
1390
1395
1400
7980
7990
8000
8010
8020
8030
8040
8050
8060
8070
8080
8090
8100
8110
8120
8130

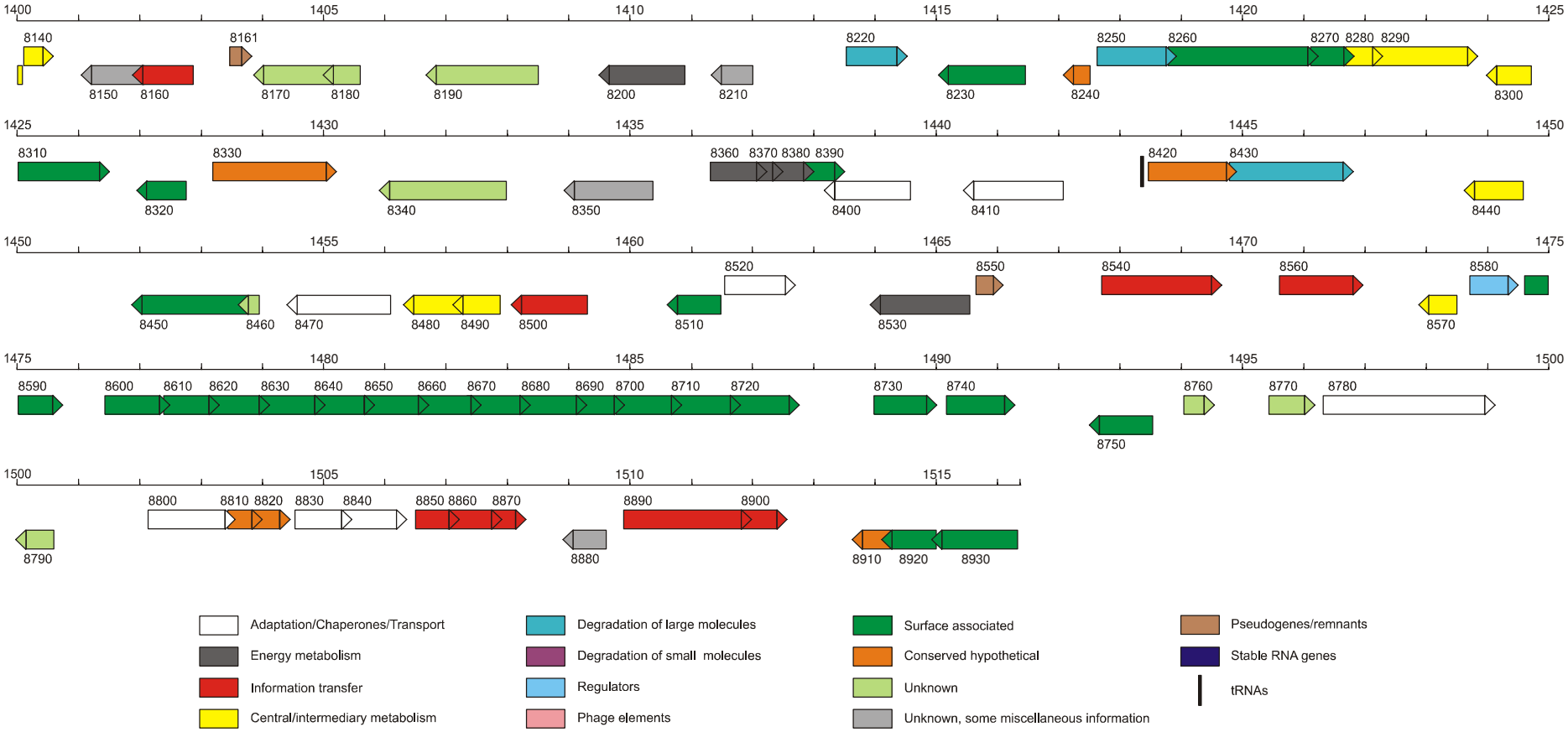


Figure 2.3. (Above and previous seven pages.) Linear representation of the *E. ruminantium* (Welgevonden isolate) genome. The scale of the map is shown in 1-kb increments. The potential protein coding regions (colour-coded by biological role) are depicted as boxes with arrowheads indicating the direction of transcription. The RNA-encoding genes are represented by dark blue boxes and the tRNAs by black bars.

2.3.2.4. Subcellular localisation of ORFs

Information on subcellular localisation is key to elucidating the functions of a protein. The proteins of Gram-negative bacteria have four major subcellular localisations: the cytoplasm, the inner membrane, the periplasm, and the outer membrane; some proteins may also be secreted extracellularly. Surface-associated proteins are of particular interest for several reasons. In many pathogenic bacteria, for instance, the invasion of host cells is mediated by surface proteins that recognize specific ligands in the extracellular matrix or on the surface of host cells (Navarre & Schneewind, 1999; Niemann *et al.,* 2004). Intracellular pathogens also rely on various membrane-associated proteins for the acquisition of metabolic intermediates, environmental signalling, cell homeostasis, and evasion of host defence systems (Finlay & Falkow, 1997; Lin *et al.,* 2002). Finally, in the case of extracellular bacteria, cell surface or secreted proteins are exposed to antibody-mediated host immune responses and are therefore primary vaccine targets (Chakravarti *et al.,* 2001).

A secreted protein is recognised by a signal peptide, a stretch of hydrophobic amino acids located at the N-terminus, and membrane proteins are characterised by one or more transmembrane helices which are similar to the signal peptide sequences. This common trait makes it difficult for signal peptide and transmembrane helix predictors to correctly assign identity to stretches of hydrophobic residues near the N-terminal methionine of a protein sequence (Yuan *et al.*, 2003). Therefore we used SignalP to identify signal sequences, TMHMM to detect transmembrane helices, and Phobius, a combined transmembrane topology and signal peptide predictor, to reduce cross-prediction errors. The results of all the searches are summarised in Appendix E.

Signal peptides were predicted for 66 CDSs, of which 13 also contained one or two predicted transmembrane helices. There are many possible membrane proteins in the *E. ruminantium* genome: 28% (247) of all CDSs, other than pseudogenes, are predicted to contain at least one transmembrane helix, 197 of which begin within the first 10 aa of the protein. Forty-eight of these transmembrane helices were also predicted to be signal sequences by the SignalP algorithm.

When compared with the results of another algorithm, Phobius, 15 of the 48 transmembrane helices were in fact predicted to be signal peptides (Appendix E) so the annotation of these CDSs is uncertain.

Two additional algorithms, pSORTb and CELLO, were utilised to assist in the assignment of proteins to subcellular localisations (Figure 2.4). However the results vary significantly between the two algorithms with only 39% of the putative proteins being assigned to the same location by each program. The majority of the shared predictions were for allocations to the cytoplasm (217 ORFs) and inner membrane (109 ORFs). Only 20 of the proteins were predicted by both algorithms to be in the outer membrane. Similar results were found by Sprenger and co-workers (2006), who compared five mammalian localisation prediction algorithms, including CELLO and WoLF PSORT (http://wolfpsort.org), and found that the different predictors generally failed to agree.

One explanation for the discrepancies in the results could be the different approaches employed by the algorithms. pSORTb does not force a prediction and will return "unknown" when a location site cannot be reliably predicted within probability limits assigned by the program, whereas CELLO designates the most likely location for each protein sequence. Of the 888 putative proteins analysed in this study 452 (51%) were returned as "unknown" by pSORTb (Figure 2.4). The CELLO results provided some indication of location for all putative proteins, even if the confidence values were low. Without experimental evidence it is not possible to determine which algorithm is the superior predictor.

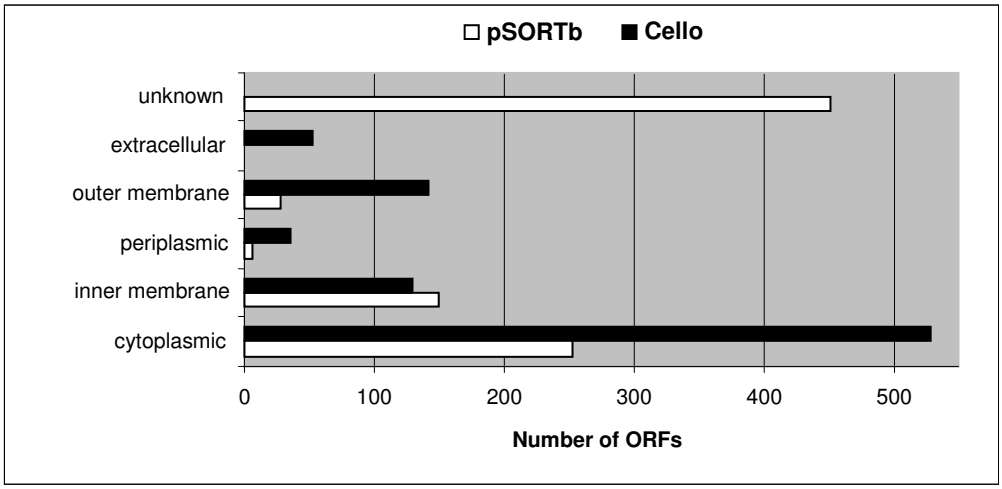


Figure 2.4. Predicted compartmentalisation of putative proteins by pSORTb and CELLO.

In addition to the limitations of the prediction programs there is the problem that it is almost impossible to predict the conditions under which proteins are expressed *in vivo*. For example, it has been shown that contact with epithelial cells leads to significant remodelling of the *Neisseria meningitidis* membrane components (Grandi, 2003). This work used DNA microarray technology to follow the changes in gene expression profiles following *N. meningitidis* interaction with human epithelial cells. Computational analysis predicted that, among the upregulated adhesion-modulated genes, only 40% of them potentially encoded inner membrane, periplasmic or outer membrane proteins. This would imply that the interaction with epithelial cells led to a change in bacterial surface protein profile, which was subsequently confirmed by fluorescent-activated cell sorting analysis. In fact, two of the proteins (glyceraldehyde 3-P dehydrogenase and N-acetylglutamate synthetase) that appeared on the surface after adhesion are predicted to be located in the cytoplasm by the available computer algorithms. While these observations do not mean that computer predictions are worthless, one should be cautious in the interpretation of prediction results. Moreover, with algorithms constantly improving (Choo *et al*., 2009; Wang & Yang, 2009; Yu *et al.*, 2010) and more experimental data becoming available, future predictions ought to be more reliable.

2.3.2.5. Paralogous gene families of membrane proteins

Several paralogous families of hypothetical membrane proteins were identified. We assigned genes to a family if they were predicted to code for proteins of similar lengths, had similar features, and had a mean of all pairwise identities that did not fall below the 15-25% ''twilight zone,'' below which a common origin is unlikely (Doolittle, 1981). Animals infected with *E. ruminantium* develop a dominant antibody response directed against an outer membrane protein, designated major antigenic protein 1 (MAP1) (Rossouw *et al*., 1990). This prominence led to *map*1 being the first *E. ruminantium* gene to be cloned and sequenced (Van Vliet *et al*., 1994), and it was subsequently found to be a member of a multigene family of outer membrane proteins which comprises 16 paralogs (Van Heerden *et al*., 2004a). Multigene families orthologous to the *map*1 family also occur in *E. canis* (Ohashi *et al*., 1998a), *E. chaffeensis* (Ohashi *et al*., 1998b), *E. muris* (Crocquet-Valdes *et al*., 2003) and *E. ewingii* (Zhang *et al*., 2008b). Recently it was shown that two proteins in the orthologous OMP family of *E. chaffeensis*, OMP19 and OMP18, function as porins that might regulate nutrient uptake during intracellular development (Kumagai *et al*., 2008).

Examination of the *E. ruminantium* genome sequence identified several other families of paralogous hypothetical membrane proteins, the two largest containing 14 and 10 paralogs respectively. The members of the first family were clustered close together, starting with Erum2240 to Erum2300 (on the reverse strand with respect to the genome numbering) followed by Erum2310 to Erum2350 on the forward strand. Two other paralogs, Erum2400 and Erum2410, separated from the rest of the family by three unrelated genes, were also on the reverse strand. All members of this family are predicted to contain either a signal peptide or a transmembrane helix close to the 5' end of the gene, and some of the latter may in fact be signal peptides. This suggests that these proteins are membrane-associated, although we do not know whether they are outer membrane constituents. The second family was located in two separate regions of the genome, with Erum2750 to Erum2800 in one cluster and Erum3600 to Erum3630 in another. Erum3600 was in the opposite orientation from the other paralogs. The members of

this family all contain a predicted signal peptide sequence and one predicted transmembrane helix and are all therefore probably outer membrane proteins. No known database homologs could be identified for the members of these two families, and for both families a BLAST search of the *E. chaffeensis* genome revealed no orthologs.

A small family of four predicted integral membrane proteins, Erum7990, Erum8000, Erum8010, Erum8020, was related to a number of hypothetical proteins in *Anaplasma marginale*, some of which have been identified as being members of the *msp*2 superfamily (ORF X, ORF Y and OMP2). In *Anaplasma marginale* MSP2 and MSP3 are immunodominant outer membrane proteins that generate antigenic diversity by recombination of variable pseudogenes, which are widely dispersed throughout the genome, into a functional expression site (Meeus *et al.*, 2003). The gene X (ORF X) multigene family is associated with *msp*2 and *msp*3 pseudogenes and may be involved in a similar mechanism for generation of antigenic variation (Meeus & Barbet, 2001). However, it is unlikely that the four *E. ruminantium* genes provide a similar variation mechanism since we could not identify any other paralogs, or orthologs of *msp*2 or *msp*3. Although *msp*2 is similar to *map*1 their arrangement within the genome is different. The *msp*2 and *msp*3 genes and pseudogenes are dispersed throughout the *A. marginale* genome, while in *E. ruminantium* families of putative outer membrane genes (including the *map*1 multigene family) appear to consist of full length genes located in tandem.

2.3.2.6. Pathogenicity-associated genes

A type IV secretion system was identified in the *E. ruminantium* genome that contains several homologs of the *virB* gene operon. There were two clusters of *virB* genes in the *E. ruminantium* genome: v*irD4*, *virB8, virB9*, *virB10* and *virB11* were grouped together, while the second locus consisted of *virB3*, *virB4*, *virB6* and three additional large genes, Erum5210, Erum5220, Erum5230, which probably encode type IV secretion proteins. Additional *virB8* and *virB4* homologs were not associated with these clusters. The *E. canis virB9* has been cloned and expressed, and was found to be highly antigenic (Felek *et al.*, 2003), it is therefore considered to

be a possible vaccine candidate for canine ehrlichiosis. Furthermore, *virB9* and *virB10* of *A. marginale* were identified in a protective outer membrane vaccine (Lopez *et al.*, 2007). The *virB1*, *virB2*, *virB5* and *virB7* genes, as well as genes encoding the proteins VirA and VirG responsible for regulating the expression of the virB locus in *Agrobacterium tumefaciens* (Thompson *et al.*, 1988; Das & Pazour, 1989) do not appear to be present in *E. ruminantium*. Genes encoding the known effector proteins VirD2, VirE2 and VirF were not found but a putative *trbG* gene, involved in conjugal transfer of T-DNA in *A. tumefaciens*, was located 388 kb away from the nearest *virB* gene clusters. Many genes which are normally clustered in operons in other bacteria are dispersed in *E. ruminantium*, so it may be significant that the normal *virB* operon structure is maintained.

The function of the type IV secretion system identified in *E. ruminantium* is unknown, but it may be involved in pathogenesis. Type IV secretion systems have been implicated as essential virulence factors in several other pathogenic bacteria. *Helicobacter pylori* uses the Cag system to deliver a 145 kDa CagA protein to mammalian cells; CagA is responsible for a number of changes in host cell physiology (Segal *et al.*, 1999) and has antiphagocytic properties (Ramarao *et al.*, 2000). *Legionella pneumophila*, *Brucella suis*, *B. abortus* and *Bartonella henselae* are thought to use type IV secretion systems to export effector proteins that contribute to survival within phagosomes (reviewed in Christie, 2001). *Bordetella pertussis* secretes pertussis toxin (PT) to the extracellular milieu using the Ptl system (Weiss *et al.*, 1993), PT itself interacts with mammalian cells rather than the type IV secretion machinery.

2.4. CONCLUSIONS

The entire genome sequence of *E. ruminantium* has been determined using a shotgun sequencing strategy. We identified 888 putative protein encoding genes and a preliminary functional analysis has identified a variety of possible surface-associated proteins and virulence factors which merit further investigation. Genome annotation is an ongoing process and requires continuous updating

of all information. Because 41% of the putative proteins are similar to hypothetical proteins of unknown function, a situation seen in other completed microbial genomes, a substantial portion of *E. ruminantium*'s biochemistry and cell biology remains to be discovered.

Homology-based annotation will often include incomplete or erroneous predictions of gene function. Just a few changes in an enzyme's active site may alter its substrate specificity, and in the absence of experimental evidence the best match does not necessarily represent a true ortholog. A metabolic function can be carried out by proteins that are completely unrelated to known enzymes, or by molecules that are so divergent that they are not regarded as homologs (Moxon *et al*., 2002). However, despite the limitations of annotation based on homology, this approach provides valuable information about the biology of the organism and provides a starting point for future experiments. The challenge now is to exploit the raw data of the genome sequence to understand the *in vivo* behaviour of the pathogen.

Table 2.2. Functional classification of *Ehrlichia ruminantium* protein-coding genes. ORF identification numbers correspond to those in Figure 2.3. The number of predicted genes in each category is indicated in brackets. (Adapted from Collins *et al.,* 2005. [Supplementary information]).

ENERGY METABOLISM (56)		
ATP-synthase complex (8)		
Erum0820	*atpA*	ATP synthase alpha chain
Erum8360	*atpB*	ATP synthase A subunit
Erum4580	*atpC*	ATP synthase epsilon chain
Erum4590	*atpD*	ATP synthase beta chain
Erum8370	*atpE*	ATP synthase C subunit
Erum8380	*atpF*	probable ATP synthase B subunit
Erum3990	*atpG*	ATP synthase gamma chain
Erum0830	*atpH*	probable ATP synthase delta chain
Electron transport (34)		
Erum7740	*coxA*	probable cytochrome c oxidase subunit I
Erum7730	*coxB*	probable cytochrome c oxidase subunit II
Erum0170	*coxC*	cytochrome c oxidase subunit III
Erum0240	*fdxA*	ferredoxin
Erum4200	*fdxB*	ferredoxin, 2FE-2S
Erum3100	*nuoA*	probable NADH-quinone oxidoreductase chain A
Erum3090	*nuoB*	NADH-quinone oxidoreductase chain B
Erum3070	*nuoC*	probable NADH-quinone oxidoreductase chain C
Erum4420	*nuoD*	NADH-quinone oxidoreductase chain D
Erum4430	*nuoE*	NADH-quinone oxidoreductase chain E
Erum4810	*nuoF*	NADH-quinone oxidoreductase chain F
Erum4270	*nuoG*	NADH-quinone oxidoreductase chain G
Erum4280	*nuoH*	NADH-quinone oxidoreductase chain H
Erum3710	*nuoI*	NADH-quinone oxidoreductase chain I
Erum4800	*nuoJ*	NADH-quinone oxidoreductase chain J
Erum4790	*nuoK*	NADH-quinone oxidoreductase chain K
Erum4780	*nuoL*	NADH-quinone oxidoreductase chain L
Erum4770	*nuoM*	NADH-quinone oxidoreductase chain M
Erum4760	*nuoN*	NADH-quinone oxidoreductase chain N
Erum5040	*petA*	ubiquinol-cytochrome c reductase iron-sulphur subunit
Erum5030	*petB*	cytochrome b
Erum5020	*petC*	cytochrome c1 precursor
Erum6260	*qor*	probable quinone oxidoreductase
Erum6810	*sdhA*	succinate dehydrogenase flavoprotein subunit
Erum6800	*sdhB*	succinate dehydrogenase iron-sulfur subunit
Erum1890	*sdhC*	probable succinate dehydrogenase cytochrome b-556 subunit
Erum1891	*sdhD*	probable succinate dehydrogenase cytochrome b small subunit
Erum0430		possible NADH-ubiquinone oxidoreductase subunit
Erum1240		probable NADH-quinone oxidoreductase subunit
Erum1570		probable cytochrome b561
Erum5440		probable NADH-quinone oxidoreductase subunit
Erum6700		probable NADH-quinone oxidoreductase subunit
Erum6720		probable c-type cytochrome
Erum7570		probable NADH-ubiquinone oxidoreductase
Pyruvate dehydrogenase and TCA cycle (14)		
Erum7920	*acnA*	aconitate hydratase
Erum6330	*fumC*	fumarate hydratase class II
Erum0750	*gltA*	citrate synthase
Erum8530	*icd*	isocitrate dehydrogenase [NADP]

Erum4090	*mdh*	malate dehydrogenase
Erum7520	*pdhA*	pyruvate dehydrogenase E1 component, alpha subunit
Erum0980	*pdhB*	probable pyruvate dehydrogenase E1 component, beta subunit
Erum0670	*pdhC*	dihydrolipoamide acetyltransferase, E2 component of pyruvate dehydrogenase complex
Erum2650	*sucA*	2-oxoglutarate dehydrogenase E1 component
Erum8200	*sucB*	dihydrolipoamide succinyltransferase, E2 component of 2-oxoglutarate dehydrogenase complex
Erum1520	*sucC*	succinyl-CoA synthetase, beta subunit
Erum1510	*sucD*	succinyl-CoA synthetase, alpha subunit
Erum1420		probable dihydrolipoamide dehydrogenase, E3 component of pyruvate or 2-oxoglutarate dehydrogenase complex
Erum5130		probable dihydrolipoamide dehydrogenase, E3 component of pyruvate or 2-oxoglutarate dehydrogenase complex
CENTRAL INTERMEDIARY METABOLISM (24)		
Erum4840	*eno*	enolase
Erum0650	*fbaB*	probable fructose-bisphosphate aldolase class I
Erum0010	*gapB*	NAD(P)-dependent glyceraldehyde 3-phosphate dehydrogenase
Erum6470	*glpX*	fructose-1,6-bisphosphatase class II GlpX
Erum5150	*gpmI*	2,3-bisphosphoglycerate-independent phosphoglycerate mutase
Erum1200	*maeB*	NADP-dependent malic enzyme
Erum0070	*metK*	S-adenosylmethionine synthetase
Erum8570	*ndk*	nucleoside diphosphate kinase
Erum0360	*pgk*	phosphoglycerate kinase
Erum7840	*ppa*	inorganic pyrophosphatase
Erum6690	*ppdK*	pyruvate phosphate dikinase
Erum7490	*ppnK*	probable inorganic polyphosphate/ATP-NAD kinase
Erum7240	*pyrH*	uridylate kinase
Erum0560	*rpe*	ribulose-phosphate 3-epimerase
Erum4100	*rpiB*	ribose 5-phosphate isomerase B
Erum4570	*tal*	probable transaldolase
Erum5600	*tkt*	transketolase
Erum4040	*tpiA*	triosephosphate isomerase
Erum0890		probable aminomethyl transferase
Erum1560		probable 2-nitropropane dioxygenase
Erum2530		probable glutathione S-transferase
Erum3230		possible NAD-glutamate dehydrogenase
Erum4020		probable pyridine nucleotide-oxidoreductase
Erum4160		probable NifU-like protein
PURINE AND PYRIMIDINE METABOLISM (29)		
Deoxyribonucleotide metabolism (3)		
Erum5190	*dut*	probable deoxyuridine 5'-triphosphate nucleotidohydrolase
Erum5650	*nrdA*	probable ribonucleoside-diphosphate reductase alpha chain
Erum3270	*nrdB*	probable ribonucleoside-diphosphate reductase beta chain
Purine ribonucleotide biosynthesis (17)		
Erum5880	*adk*	adenylate kinase
Erum6740	*gmk*	guanylate kinase
Erum0740	*guaA*	GMP synthase [glutamine-hydrolyzing]
Erum7500	*guaB*	inosine-5'-monophosphate dehydrogenase
Erum7900	*prsA*	ribose-phosphate pyrophosphokinase
Erum5630	*purA*	adenylosuccinate synthetase
Erum2460	*purB*	adenylosuccinate lyase
Erum7000	*purC*	phosphoribosylaminoimidazole-succinocarboxamide synthase
Erum7770	*purD*	phosphoribosylamine--glycine ligase
Erum1060	*purE*	phosphoribosylaminoimidazole carboxylase catalytic subunit
Erum0900	*purF*	glutamine phosphoribosylpyrophosphate amidotransferase
Erum8290	*purH*	bifunctional purine biosynthesis protein PurH
Erum7940	*purK*	phosphoribosylaminoimidazole carboxylase ATPase subunit
Erum6510	*purL*	probable phosphoribosylformylglycinamidine synthase II

Erum6580	*purM*	phosphoribosylformylglycinamidine cyclo-ligase
Erum6370	*purN*	phosphoribosylglycinamide formyltransferase
Erum6450	*purQ*	possible phosphoribosylformylglycinamidine synthase I
Pyrimidine ribonucleotide biosynthesis (9)		
Erum6110	*cmk*	probable kinase
Erum6990	*dcd*	probable deoxycytidine triphosphate deaminase
Erum4250	*pyrB*	aspartate carbamoyltransferase
Erum6350	*pyrC*	dihydroorotase
Erum1810	*pyrD*	dihydroorotate dehydrogenase
Erum8490	*pyrE*	probable phosphoribosyltransferase
Erum3040	*pyrF*	orotidine 5'-phosphate decarboxylase
Erum1160	*pyrG*	CTP synthase
Erum7460	*tmk*	probable thymidylate kinase
FATTY ACID METABOLISM (12)		
Erum3430	*acpS*	probable holo-[acyl-carrier-protein] synthase
Erum5320	*bccA*	probable acetyl-/propionyl-coenzyme A carboxylase alpha chain
Erum7470	*fabD*	probable malonyl CoA-acyl carrier protein transacylase
Erum2150	*fabF*	3-oxoacyl-[acyl-carrier-protein] synthase II
Erum3840	*fabG*	3-oxoacyl-[acyl carrier protein] reductase
Erum5720	*fabH*	3-oxoacyl-[acyl-carrier-protein] synthase III
Erum2860	*fabI*	enoyl-[acyl-carrier-protein] reductase [NADH]
Erum8280	*fabZ*	(3R)-hydroxymyristoyl-[acyl carrier protein] dehydratase
Erum2840	*matA*	probable malonyl-CoA decarboxylase
Erum0550	*plsC*	probable 1-acyl-sn-glycerol-3-phosphate acyltransferase
Erum5730	*plsX*	fatty acid/phospholipid synthesis protein
Erum7220		probable cytidylyltransferase
MACROMOLECULE SYNTHESIS AND MODIFICATION (19)		
Erum3060	*ccmE*	cytochrome c-type biogenesis protein CcmE
Erum7750	*ctaB*	probable protoheme IX farnesyltransferase
Erum8080	*ctaG*	cytochrome c oxidase assembly protein
Erum0880	*ccmF*	cytochrome c-type biogenesis protein CcmF
Erum2210	*dsbB*	disulfide bond formation protein B
Erum6910	*dsbE*	probable thiol:disulfide interchange protein
Erum6600	*gpsA*	glycerol-3-phosphate dehydrogenase [NAD(P)+]
Erum8440	*lgt*	prolipoprotein diacylglyceryl transferase
Erum6360	*lipB*	lipoate-protein ligase B
Erum1220	*lnt*	probable apolipoprotein N-acyltransferase
Erum8120	*lspA*	lipoprotein signal peptidase
Erum3370	*mdmC*	probable O-methyltransferase
Erum1980	*pgpA*	probable phosphatidylglycerophosphatase A
Erum8300	*pgsA*	probable CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase
Erum3160	*pssA*	probable CDP-diacylglycerol--serine O-phosphatidyltransferase
Erum3170	*psd*	probable phosphatidylserine decarboxylase proenzyme
Erum3720	*sipF*	prokaryotic type I signal peptidase
Erum4211		possible cytochrome c-type biogenesis protein
Erum7040		probable cytochrome c oxidase assembly protein
AMINO ACID METABOLISM (26)		
Erum3490	*aatA*	aspartate aminotransferase A
Erum4480	*argB*	acetylglutamate kinase
Erum7830	*argC*	N-acetyl-gamma-glutamyl-phosphate reductase
Erum2110	*argD*	acetylornithine/succinyldiaminopimelate aminotransferase
Erum0510	*argF*	ornithine carbamoyltransferase
Erum3770	*argG*	argininosuccinate synthase
Erum1830	*argH*	argininosuccinate lyase
Erum3800	*argJ*	arginine biosynthesis bifunctional protein ArgJ
Erum0060	*asd*	aspartate-semialdehyde dehydrogenase
Erum1880	*aroE*	3-phosphoshikimate 1-carboxyvinyltransferase
Erum5170	*carA*	carbamoyl-phosphate synthase small chain

Erum6310	*carB*	carbamoyl-phosphate synthase, large subunit
Erum2670	*dapA*	dihydrodipicolinate synthase
Erum5770	*dapB*	dihydrodipicolinate reductase
Erum0390	*dapD*	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase
Erum0940	*dapE*	probable succinyl-diaminopimelate desuccinylase
Erum0340	*dapF*	diaminopimelate epimerase
Erum0610	*glnA*	glutamine synthetase
Erum6840	*glyA*	serine hydroxymethyltransferase
Erum4150	*iscS*	cysteine desulfurase
Erum5340	*lysA*	probable diaminopimelate decarboxylase
Erum4460	*pccB*	propionyl-CoA carboxylase beta chain
Erum0030	*proC*	pyrroline-5-carboxylate reductase
Erum3850	*putA*	proline dehydrogenase/delta-1-pyrroline-5-carboxylate dehydrogenase
Erum1480		possible truncated glutamine synthetase
Erum7720		probable aspartate kinase
BIOSYNTHESIS OF CO-FACTORS (61)		
Biotin biosynthesis (5)		
Erum3870	*bioA*	adenosylmethionine-8-amino-7-oxononanoate aminotransferase
Erum6500	*bioB*	biotin synthase
Erum0220	*bioC*	possible biotin synthesis protein BioC
Erum1740	*bioF*	probable 8-amino-7-oxononanoate synthase
Erum2520		probable biotin--[acetyl-CoA-carboxylase] synthetase
Folic acid (7)		
Erum4080	*folB*	possible dihydroneopterin aldolase
Erum3680	*folC*	probable folylpolyglutamate synthase/dihydrofolate synthase
Erum6730	*folD*	methylenetetrahydrofolate dehydrogenase/ methenyltetrahydrofolate cyclohydrolase
Erum4000	*folE*	GTP cyclohydrolase I
Erum6520	*folK*	probable 2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase
Erum6280	*folP1*	probable dihydropteroate synthase 1
Erum6290	*folP2*	probable dihydropteroate synthase 2
Heme and porphyrins (7)		
Erum0630	*hemA*	5-aminolevulinic acid synthase
Erum2720	*hemB*	delta-aminolevulinic acid dehydratase
Erum3690	*hemC*	porphobilinogen deaminase
Erum5380	*hemD*	probable uroporphyrinogen-III synthase
Erum0180	*hemE*	uroporphyrinogen decarboxylase
Erum4550	*hemF*	coproporphyrinogen III oxidase
Erum6180	*hemH*	ferrochelatase
Menaquinone and ubiquinones (13)		
Erum4750	*dxr*	1-deoxy-D-xylulose 5-phosphate reductoisomerase
Erum5660	*ispA*	probable geranyltranstransferase
Erum0600	*ispB*	octaprenyl-diphosphate synthase
Erum1030	*ispD*	probable 2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase
Erum3340	*ispE*	probable 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase
Erum1020	*ispF*	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase
Erum4730	*ispG*	probable 1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase
Erum5180	*ispH*	4-hydroxy-3-methylbut-2-enyl diphosphate reductase
Erum5790	*ubiA*	4-hydroxybenzoate octaprenyltransferase
Erum2600	*ubiB*	probable ubiquinone biosynthesis protein UbiB
Erum7700	*ubiE*	ubiquinone/menaquinone biosynthesis methyltransferase UbiE
Erum0080	*ubiF*	probable 2-octaprenyl-3-methyl-6-methoxy-1,4-benzoquinol hydroxylase
Erum4110	*ubiG*	probable 3-demethylubiquinone-9 3-methyltransferase
Riboflavin (6)		
Erum0800	*ribB*	3,4-dihydroxy-2-butanone 4-phosphate synthase
Erum7390	*ribE*	probable riboflavin synthase, alpha subunit
Erum1140	*ribD*	riboflavin biosynthesis protein RibD
Erum8130	*ribF*	riboflavin kinase/FAD synthetase

Erum3130	*ribH*	probable 6,7-dimethyl-8-ribityllumazine synthase
Erum0310		probable riboflavin biosynthesis protein
Thiamine (8)		
Erum2970	*thiC*	thiamine biosynthesis protein ThiC
Erum1910	*thiD*	probable phosphomethylpyrimidine kinase
Erum2060	*thiE*	probable thiamine-phosphate pyrophosphorylase
Erum8480	*thiF*	probable adenylyltransferase ThiF
Erum7630	*thiG*	thiazole biosynthesis protein
Erum4980	*thiL*	probable thiamine-monophosphate kinase
Erum5680	*thiO*	probable thiamine biosynthesis oxidoreductase
Erum7640		thiamin S protein
Other (15)		
Erum2160	*acpP*	acyl carrier protein
Erum2960	*coaE*	probable dephospho-CoA kinase
Erum3460	*coaD*	probable phosphopantetheine adenylyltransferase
Erum8140	*grxC*	probable glutaredoxin 3
Erum0770	*gshA*	possible gamma-glutamylcysteine synthetase
Erum6640	*gshB*	glutathione synthetase
Erum5290	*lipA*	lipoic acid synthetase
Erum0230	*nadA*	quinolinate synthetase A
Erum0140	*nadC*	nicotinate-nucleotide pyrophosphorylase [carboxylating]
Erum2910	*nadD*	probable nicotinate-nucleotide adenylyltransferase
Erum2710	*nadE*	probable glutamine-dependent NAD(+) synthetase
Erum1850	*pdxH*	pyridoxamine 5'-phosphatate oxidase
Erum2920	*pdxJ*	pyridoxal phosphate biosynthetic protein PdxJ
Erum7540	*trxA*	thioredoxin 1
Erum3470	*trxB*	thioredoxin reductase
INFORMATION TRANSFER (173)		
DNA replication, repair, recombination and degradation (48)		
Erum0410	*dfp*	probable DNA/pantothenate metabolism flavoprotein
Erum2870	*dnaA*	chromosomal replication initiator protein DnaA
Erum5710	*dnaB*	replicative DNA helicase
Erum1870	*dnaE*	DNA polymerase III, alpha subunit
Erum3310	*dnaG*	probable DNA primase
Erum7880	*dnaN*	DNA polymerase III, beta subunit
Erum4990	*dnaQ*	DNA polymerase III, epsilon subunit
Erum0040	*dnaZ*	probable DNA polymerase III, gamma subunit
Erum3810	*exoA*	probable exodeoxyribonuclease
Erum2420	*gyrA*	DNA gyrase subunit A
Erum4260	*gyrB*	DNA gyrase subunit B
Erum2940	*holB*	DNA III, delta' subunit
Erum2930	*hupB*	DNA-binding protein HU-beta
Erum1080	*ihfA*	probable integration host factor alpha subunit
Erum6140	*ihfB*	possible integration host factor beta subunit
Erum6940	*ligA*	NAD-dependent DNA ligase
Erum7290	*mfd*	transcription-repair coupling factor
Erum2130	*mutL*	DNA mismatch repair protein MutL
Erum4330	*mutM*	formamidopyrimidine-DNA glycosylase
Erum2700	*mutS*	DNA mismatch repair protein MutS
Erum2430	*nth*	endonuclease III
Erum0490	*polA*	DNA polymerase I
Erum5360	*priA*	primosomal protein N'
Erum6900	*radA*	DNA repair protein RadA
Erum6440	*radC*	DNA repair protein RadC
Erum8500	*recA*	RecA protein (Recombinase A)
Erum6250	*recB*	probable exodeoxyribonuclease V beta chain
Erum0520	*recF*	probable DNA replication and repair protein RecF
Erum0420	*recG*	ATP-dependent DNA helicase RecG

Erum8550	*recJ*	probable single-stranded-DNA-specific exonuclease RecJ
Erum4920	*recO*	possible DNA repair protein RecO
Erum2570	*recR*	probable recombination protein RecR
Erum4520	*rmuC*	DNA recombination protein RmuC
Erum6760	*ruvA*	probable junction DNA helicase RuvA
Erum6770	*ruvB*	Holliday junction DNA helicase RuvB
Erum0160	*ruvC*	crossover junction endodeoxyribonuclease RuvC
Erum2140	*smf*	DNA processing protein chain A
Erum2830	*ssb*	single-strand DNA binding protein
Erum3400	*topA*	DNA topoisomerase I
Erum3110	*uvrA*	uvrABC system protein A
Erum2390	*uvrD*	DNA helicase II
Erum0370	*xseA*	exodeoxyribonuclease VII
Erum7560	*xseB*	probable exodeoxyribonuclease VII small subunit
Erum0530		possible uracil DNA glycosylase
Erum1180		probable integrase/recombinase XerD or XerC
Erum5640		possible Holliday junction resolvase
Erum6590		probable integrase/recombinase XerD or XerC
Erum7170		probable methylpurine-DNA glycosylase
Degradation of RNA (6)		
Erum3540	*pnp*	polyribonucleotide nucleotidyltransferase
Erum8070	*rnc*	ribonuclease III
Erum7260	*rnhA*	ribonuclease HI
Erum1760	*rnhB*	ribonuclease HII
Erum5800	*rnpA*	probable ribonuclease P protein component
Erum5510		probable ribonuclease
RNA synthesis and modification (12)		
Erum0810	*greA*	transcription elongation factor GreA
Erum4700	*nusA*	N utilization substance protein A
Erum1670	*nusG*	transcription antitermination protein NusG
Erum1400	*rho1*	transcription termination factor 1
Erum7670	*rho2*	transcription termination factor 2
Erum5850	*rpoA*	DNA-directed RNA polymerase alpha chain
Erum1720	*rpoB*	DNA-directed RNA polymerase beta chain
Erum1730	*rpoC*	DNA-directed RNA polymerase beta' chain
Erum3320	*rpoD*	RNA polymerase sigma-70 factor
Erum3960	*rpoH*	RNA polymerase sigma-32 factor
Erum2990	*rpoZ*	DNA-directed RNA polymerase omega chain
Erum8560		probable nucleic acid independent RNA polymerase
Aminoacyl-tRNA synthetases (21)		
Erum1500	*alaS*	alanyl-tRNA synthetase
Erum4910	*argS*	arginyl-tRNA synthetase
Erum6660	*aspS*	aspartyl-tRNA synthetase
Erum3250	*cysS*	cysteinyl-tRNA synthetase
Erum7610	*gltX1*	glutamyl-tRNA synthetase 1
Erum4310	*gltX2*	glutamyl-tRNA synthetase 2
Erum0110	*glyQ*	glycyl-tRNA synthetase alpha chain
Erum0120	*glyS*	glycyl-tRNA synthetase beta chain
Erum7010	*hisS*	histidyl-tRNA synthetase
Erum4870	*ileS*	isoleucyl-tRNA synthetase
Erum3010	*leuS*	leucyl-tRNA synthetase
Erum4220	*lysS*	lysyl-tRNA synthetase
Erum7710	*metG*	methionyl-tRNA synthetase
Erum1360	*pheS*	phenylalanyl-tRNA synthetase alpha chain
Erum5830	*pheT*	phenylalanyl-tRNA synthetase beta chain
Erum3440	*proS*	prolyl-tRNA synthetase
Erum4540	*serS*	seryl-tRNA synthetase
Erum8890	*thrS*	threonyl-tRNA synthetase
Erum1120	*trpS*	tryptophanyl-tRNA synthetase

Erum0620	*tyrS*	tyrosyl-tRNA synthetase
Erum0780	*valS*	valyl-tRNA synthetase
tRNA and aminoacyl-tRNA modification (17)		
Erum0540	*def1*	probable deformylase 1
Erum1820	*def2*	probable peptide deformylase 2
Erum2030	*fmt*	methionyl-tRNA formyltransferase
Erum3670	*gatA*	glutamyl-tRNA(Gln) amidotransferase subunit A
Erum2850	*gatB*	aspartyl/glutamyl-tRNA amidotransferase subunit B
Erum7910	*gatC*	probable glutamyl-tRNA(Gln) amidotransferase subunit C
Erum4030	*ksgA*	dimethyladenosine transferase
Erum4370	*miaA*	probable tRNA delta(2)-isopentenylpyrophoshate transferase
Erum0910	*pth*	peptidyl-tRNA hydrolase
Erum4970	*rbn*	tRNA processing ribonuclease BN
Erum5750	*tgt*	queuine tRNA-ribosyltransferase
Erum8860	*trmD*	tRNA (Guanine-N(1)-)-methyltransferase
Erum0400	*trmE*	probable tRNA modification GTPase
Erum2230	*trmU*	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase
Erum4240	*truA*	tRNA pseudouridine synthase A
Erum3520	*truB*	probable tRNA pseudouridine synthase B
Erum6100		probable tRNA/rRNA methyltransferase
Translation factors, modification of ribosomes and nascent peptides (16)		
Erum3190	*efp*	probable elongation factor P
Erum7230	*frr*	ribosome recycling factor
Erum1650	*fusA*	elongation factor G
Erum5110	*infA*	translation initiation factor IF-1
Erum4690	*infB*	translation initiation factor IF-2
Erum8900	*infC*	translation initiation factor IF-3
Erum4500	*prfA*	peptide chain release factor 1
Erum3650	*prfB*	peptide chain release factor 2
Erum4680	*rbfA*	ribosome-binding factor A
Erum8850	*rimM*	probable 16S rRNA processing protein
Erum3210	*rluC*	ribosomal large subunit pseudouridine synthase C
Erum5330	*rluD*	ribosomal large subunit pseudouridine synthase D
Erum0790	*smpB*	SsrA-binding protein
Erum5080	*tsf*	elongation factor Ts
Erum1660	*tufA*	elongation factor Tu-A
Erum6090	*tufB*	elongation factor Tu-B
Ribosomal proteins (53)		
Erum1690	*rplA*	50S ribosomal protein L1
Erum6040	*rplB*	50S ribosomal protein L2
Erum6070	*rplC*	50S ribosomal protein L3
Erum6060	*rplD*	50S ribosomal protein L4
Erum5960	*rplE*	50S ribosomal protein L5
Erum5930	*rplF*	50S ribosomal protein L6
Erum6850	*rplI*	50S ribosomal protein L9
Erum1700	*rplJ*	50S ribosomal protein L10
Erum1680	*rplK*	50S ribosomal protein L11
Erum1710	*rplL*	50S ribosomal protein L7/L12
Erum7810	*rplM*	50S ribosomal protein L13
Erum5980	*rplN*	50S ribosomal protein L14
Erum5900	*rplO*	50S ribosomal protein L15
Erum6000	*rplP*	50S ribosomal protein L16
Erum5840	*rplQ*	50S ribosomal protein L17
Erum5920	*rplR*	50S ribosomal protein L18
Erum8870	*rplS*	50S ribosomal protein L19
Erum1370	*rplT*	50S ribosomal protein L20
Erum4830	*rplU*	50S ribosomal protein L21
Erum6020	*rplV*	50S ribosomal protein L22
Erum6050	*rplW*	50S ribosomal protein L23

Erum5970	*rplX*	50S ribosomal protein L24
Erum0920	*rplY*	probable 50S ribosomal protein L25
Erum4820	*rpmA*	50S ribosomal protein L27
Erum5350	*rpmB*	50S ribosomal protein L28
Erum5991	*rpmC*	50S ribosomal protein L29
Erum7480	*rpmE*	50S ribosomal protein L31
Erum5740	*rpmF*	50S ribosomal protein L32
Erum2190	*rpmG*	50S ribosomal protein L33
Erum5791	*rpmH*	50S ribosomal protein L34
Erum1380	*rpmI*	50S ribosomal protein L35
Erum3950	*rpmJ*	50S ribosomal protein L36
Erum6120	*rpsA*	30S ribosomal protein S1
Erum5090	*rpsB*	30S ribosomal protein S2
Erum6010	*rpsC*	30S ribosomal protein S3
Erum1940	*rpsD*	30S ribosomal protein S4
Erum5910	*rpsE*	30S ribosomal protein S5
Erum6870	*rpsF*	30S ribosomal protein S6
Erum1640	*rpsG*	30S ribosomal protein S7
Erum5940	*rpsH*	30S ribosomal protein S8
Erum7820	*rpsI*	30S ribosomal protein S9
Erum6080	*rpsJ*	30S ribosomal protein S10
Erum5860	*rpsK*	30S ribosomal protein S11
Erum1630	*rpsL*	30S ribosomal protein S12
Erum5870	*rpsM*	30S ribosomal protein S13
Erum5950	*rpsN*	30S ribosomal protein S14
Erum3530	*rpsO*	30S ribosomal protein S15
Erum1320	*rpsP*	30S ribosomal protein S16
Erum5990	*rpsQ*	30S ribosomal protein S17
Erum6860	*rpsR*	30S ribosomal protein S18
Erum6030	*rpsS*	30S ribosomal protein S19
Erum0480	*rpsT*	30S ribosomal protein S20
Erum1530	*rpsU*	possible 30S ribosomal protein S21
DEGRADATION OF PROTEINS (18)		
Erum4660	*clpA*	ATP-dependent Clp protease, ATP-binding subunit
Erum2000	*clpP*	ATP-dependent Clp protease proteolytic subunit
Erum2010	*clpX*	ATP-dependent Clp protease ATP-binding subunit ClpX
Erum4060	*gcp*	o-sialoglycoprotein endopeptidase
Erum7680	*hslV*	ATP-dependent protease HslV
Erum7690	*hslU*	ATP-dependent hsl protease ATP-binding subunit
Erum2020	*lon*	ATP-dependent protease La
Erum8160	*map*	methionine aminopeptidase
Erum6380	*pepA*	cytosol aminopeptidase
Erum3510		possible glycoprotease
Erum5610		possible carboxypeptidase
Erum6130		probable peptidase
Erum7410		probable zinc protease
Erum8050		probable exported serine protease
Erum8090		probable exported peptidase
Erum8100		probable exported M16 family peptidase
Erum8220		probable exported D-alanyl-D-alanine carboxypeptidase
Erum8250		probable membrane-associated zinc metalloprotease
CELL PROCESSES (27)		
Cell division (8)		
Erum4490	*engB*	probable GTP protein EngB
Erum8400	*ftsA*	cell division protein FtsA
Erum8430	*ftsH*	cell division protein FtsH
Erum2090	*ftsK*	probable cell division protein FtsK
Erum6620	*ftsQ*	probable cell division protein FtsQ
Erum8520	*ftsY*	probable cell division protein FtsY
Erum8800	*ftsZ*	cell division protein FtsZ
Erum6460	*gidA*	glucose inhibited division protein A

Chromosome replication (2)		
Erum8830	*parA*	chromosome partitioning protein ParA
Erum8840	*parB*	chromosome partitioning protein ParB
Chaperones (12)		
Erum6400	*clpB*	heat shock protein ClpB
Erum0130	*dnaJ*	chaperone protein DnaJ
Erum5500	*dnaK*	chaperone protein DnaK
Erum6420	*groEL*	60 kDa chaperonin GroEL
Erum6430	*groES*	10 kDa chaperonin GroES
Erum1130	*grpE*	GrpE protein
Erum4180	*hscB*	possible co-chaperone protein HscB
Erum4190	*hscA*	chaperone protein HscA
Erum2450	*htpG*	chaperone protein HtpG
Erum4010	*pmbA*	probable PmbA protein
Erum3500	*ppiD*	probable peptidyl-prolyl cis-trans isomerase D
Erum7030		probable disulfide oxidoreductase
Adaptation to atypical conditions (5)		
Erum3350	*cutA*	probable periplasmic divalent cation tolerance protein CutA
Erum0440	*dksA*	probable DnaK suppressor protein
Erum3050	*surE*	acid phosphatase SurE
Erum5270	*sodB*	superoxide dismutase [Fe]
Erum3480		probable peroxiredoxin
PATHOGENICITY-ASSOCIATED GENES (14)		
Erum5260	*virB3*	type IV secretion system protein VirB3
Erum5250	*virB4*	type IV secretion system protein VirB4
Erum5240	*virB6*	type IV secretion system protein VirB6
Erum0300	*virB8*	type IV secretion system protein VirB8
Erum0290	*virB9*	type IV secretion system protein VirB9
Erum0280	*virB10*	type IV secretion system protein VirB10
Erum0270	*virB11*	type IV secretion system protein VirB11
Erum0260	*virD4*	type IV secretion system protein VirD4
Erum4410		possible type IV secretion system protein
Erum5210		possible type IV secretion system protein
Erum5220		possible type IV secretion system protein
Erum5230		possible type IV secretion system protein
Erum7530		probable conjugal transfer protein
Erum7980		possible type IV secretion system protein
TRANSPORTERS (49)		
ABC transporters (16)		
Erum7050	*ccmA*	heme exporter protein A
Erum0450	*ccmB*	possible heme exporter protein B
Erum6750	*ccmC*	heme exporter protein C
Erum1190	*lolD*	lipoprotein releasing system ATP-binding protein LolD
Erum0860	*lolE*	probable lipoprotein releasing system transmembrane protein LolE
Erum5760	*pstB*	probable phosphate ABC transporter, ATP-binding protein
Erum0580		probable ABC transporter, ATP binding protein
Erum1490		possible ABC transporter, membrane-spanning protein
Erum1580		probable ABC transporter, membrane-spanning protein
Erum2550		probable ABC transporter, ATP-binding protein
Erum2580		probable ABC transporter, periplasmic solute binding protein
Erum2590		probable ABC transporter, ATP-binding protein
Erum5060		probable ABC transporter, membrane-spanning protein
Erum5280		probable ABC transporter, membrane-spanning protein
Erum6270		probable ABC transporter, ATP-binding protein
Erum6820		probable ABC transporter, ATP-binding and membrane-spanning protein
Amino acids (2)		
Erum1130	*proP*	proline/betaine transporter
Erum4510		probable sodium:dicarboxylate symporter(glutamate)
Proteins and peptides (11)		
Erum5430	*ffh*	signal recognition particle protein
Erum8780	*secA*	preprotein translocase SecAsubunit

Erum7430	*secB*	probable protein-export protein SecB
Erum8470	*secD*	probable protein-export membrane protein SecD
Erum0640	*secF*	protein-export membrane protein SecF
Erum1170	*secG*	probable protein-exportmembrane protein SecG
Erum5890	*secY*	preprotein translocase secY subunit
Erum2560	*tatA*	possible Sec-independent protein translocase membrane protein
Erum4720	*tatC*	Sec-independent protein translocase protein TatC
Erum1990	*tig*	trigger factor
Erum7780		probable preprotein translocase subunit YajC
Cations (9)		
Erum0190	*corC*	possible magnesium and cobalt efflux protein
Erum1310	*fbpA*	probable iron-binding periplasmic protein
Erum8410	*trkH*	Trk system potassium uptakeprotein
Erum0460		probable cation efflux system protein
Erum0950		probable glutathione-regulated potassium-efflux system protein
Erum1780		possible Na+/H+ antiporter subunit
Erum4600		probable magnesium transporter
Erum5530		probable Na+/H+ antiporter subunit
Erum5550		probable Na+/H+ antiporter subunit
Other (11)		
Erum6780	*bcr*	probable bicyclomycin resistance protein
Erum1590		probable secretion protein
Erum2740		probable integral membrane transport protein
Erum2810		probable integral membrane transport protein
Erum2820		probable integral membrane transport protein
Erum3150		probable integral membrane transport protein
Erum4710		probable integral membrane transport protein
Erum5810		probable integral membrane transport protein
Erum5820		possible competence protein
Erum7580		probable integral membrane transport protein
Erum7800		probable outer membrane efflux protein
REGULATORY FUNCTIONS (9)		
Erum3200	*suhB*	probable inositol-1-monophosphatase
Erum1000	*tldD*	TldD protein
Erum2120		possible histidine kinase sensor component of a two-component regulatory system
Erum3220		possible response regulator component of a two-component regulatory system
Erum3360		probable two component sensor kinase
Erum6610		probable response regulator component of a two-component regulatory system
Erum6960		probable histidine kinase sensor component of a two-component regulatory system
Erum7860		probable response regulator component of a two-component regulatory system
Erum8580		possible transcriptional regulator
PHAGE RELATED (3)		
Erum0200		possible protease
Erum0210		possible genetic exchange protein
Erum2660		unknown
MEMBRANE-ASSOCIATED PROTEINS (175)		
CONSERVED HYPOTHETICAL PROTEINS (50)		
SOME MISCELLANEOUS INFORMATION, BUT NO FUNCTIONAL CLASSIFICATION (63)		
NO SIMILARITY, NO FUNCTIONAL INFORMATION (80)		

CHAPTER 3

Metabolic reconstruction and comparative genomic analysis of species within the order Rickettsiales

3.1. INTRODUCTION

The order Rickettsiales lies within the phylum Proteobacteria, class Alphaproteobacteria, and its members are intracellular bacteria which have a range of mutualistic, commensal and parasitic relationships with a taxonomically diverse set of host and vector species (Table 3.1) (Dumler *et al*., 2001; Gupta & Mok, 2007; Williams *et al*., 2007). Most of the genera in the Rickettsiales contain species that are pathogenic to animals and/or humans and the order is composed of three families, Rickettsiaceae, Anaplasmataceae and Holosporaceae (Ludwig & Klenk, 2001; Fredricks, 2006). The first member of the order to be sequenced was *Rickettsia prowazekii* (Andersson *et al*., 1998). Since then the genome sequences of numerous species of both the Rickettsiaceae and Anaplasmataceae families have been determined.

The family Anaplasmataceae consists of the genera *Anaplasma*, *Ehrlichia*, *Wolbachia* and *Neorickettsia* (Dumler *et al*., 2001). *Ehrlichia* species are intracellular tick-borne pathogens that induce flu-like symptoms in both animals and humans and the bacterial populations are maintained by tick transmission within and between wild and domestic animal populations. The genome sequences of three *Ehrlichia* species were included in this study, namely *E. ruminantium* strain Welgevonden (Collins *et al*., 2005), reported on in this thesis, *E. chaffeensis* strain Arkansas (Hotopp *et al*., 2006) and *E. canis* strain Jake (Mavromatis *et al*., 2006). *E. chaffeensis* causes monocytic ehrlichiosis, a systemic human disease in the South-Central and South-eastern United States of America, while *E. canis* infects wild and domestic canids and causes canine monocytic ehrlichiosis. The two other *E. ruminantium* genome sequences which are available (section 2.1) were not included in the current analysis since very extensive comparisons of the three *E. ruminantium* sequences have already been performed (Frutos *et al*., 2006, 2007). The

current analysis concentrates on attempting to elucidate differences in biology between the different species in the order Rickettsiales.

Within the Anaplasmataceae, *Anaplasma* and *Ehrlichia* are the two most closely related genera, and two *Anaplasma* genome sequences are available: *A. marginale* strain St. Maries (Brayton *et al*., 2005) and *A. phagocytophilum* HZ (Hotopp *et al*., 2006). *A. marginale* is the most prevalent tick-borne pathogen of cattle worldwide.

Wolbachia is one of the most abundant bacterial endosymbionts and, unlike other genera in the Anaplasmataceae, no pathogenic species have yet been identified. In their host arthropods, *Wolbachia* manipulate the host's reproductive system to ensure effective transmission to the next generation. Two *Wolbachia* genome sequences have been published, those of a *Wolbachia* endosymbiont of *Drosophila melanogaster* (*W. pipientis w*Mel) (Wu *et al*., 2004) and a *Wolbachia* endosymbiont, strain TRS, of *Brugia malayi* (*W. pipientis w*Bm) (Foster *et al*., 2005).

Neorickettsia sennetsu strain Miyayama (Hotopp *et al*., 2006) was the first species in the genus *Neorickettsia* for which the genome sequence was determined. *N. sennetsu* is a monocytotropic species that causes Sennetsu fever (previously Sennetsu ehrlichiosis) in humans. The *N. risticii* genome has also been completed recently (Lin *et al*., 2009).

In the family Rickettsiaceae we find the genera *Rickettsia* and *Orientia*. Several *Rickettsia* genomes have been sequenced, including *R. prowazekii* strain Madrid E (Andersson *et al*., 1998) from the typhus group, *R. conorii* strain Malish 7 (Ogata *et al*., 2000) and *R. felis* URRWXCal2 (Ogata *et al*., 2005) from the spotted fever group, and the non-pathogenic *R. bellii* RML369-C (Ogata *et al*., 2006). *R. felis* is the only member in the order Rickettsiales that carries plasmids and this is the first putative conjugative plasmid identified among obligate intracellular bacteria (Ogata *et al*., 2005).

Pelagibacter ubique (*Candidatus Pelagibacter ubique* HTCC1062) is a free-living oceanic bacterium which is phylogenetically classified in the order Rickettsiales based on its 16S rRNA sequence (Giovannoni *et al*., 2005). Williams and colleagues confirmed this phylogeny by using the sequences of 104 selected protein families (Williams *et al*., 2007). *P. ubique* has the smallest genome, and contains the smallest number of predicted open reading frames, of all known free-living microorganisms. However *P. ubique* is very different from all other species of Rickettsiales, it does not share the intracellular lifestyle and five out of nine proteins found in almost all α-proteobacteria except the Rickettsiales are present in *P. ubique* (Gupta & Mok, 2007). It seems that *P. ubique* diverged from all other Rickettsiales even before the common ancestor of eukaryotic mitochondria (Williams *et al*., 2007), and subsequent evolution has streamlined the genome down to the minimum required for efficient growth in an environment containing limiting amounts of nutrients.

This chapter reports on the analysis of the metabolic pathways of *E. ruminantium* and *in silico* comparison with other genome sequences in the order Rickettsiales. The twelve organisms chosen for the comparative studies are those for which complete genome sequences were published at the time this study commenced (Table 3.1, Figure 3.1), although several other annotated Rickettsiales genomes have been reported subsequently. This analysis does not attempt the huge task of comparing all the pathways in detail, although others have done so for a few selected disease-causing Rickettsiales (Hotopp *et al*., 2006; Min *et al*., 2008).

Table 3.1. Characteristics of the Rickettsiales for which genome sequences were available at the time this study commenced.

Family	Species	Vertebrate Host	Invertebrate Host	Disease Caused
Anaplasmataceae	*Ehrlichia ruminantium*	Wild and domestic ruminants	Ticks	Heartwater
	Ehrlichia canis	Wild and domestic canids	Ticks	Canine monocytic ehrlichiosis
	Ehrlichia chaffeensis	Humans, deer, dogs	Ticks	Human monocytic ehrlichiosis
	Anaplasma marginale	Cattle	Ticks	Bovine anaplasmosis
	Anaplasma phagocytophilum	Humans, deer, rodents, cats, sheep, cattle, horses, llamas, bison	Ticks	Human granulocytic anaplasmosis
	Neorickettsia sennetsu	Humans	Trematodes	Sennetsu fever
	*Wolbachia pipientis w*Mel	None	Insects	None
	*Wolbachia pipientis w*Bm	None	Filarial nematodes	None
Rickettsiaceae	*Rickettsia bellii*	None	Ticks	None
	Rickettsia conorii	Humans, rodents	Ticks	Mediterranean spotted fever
	Rickettsia felis	Cats, humans	Fleas	Spotted fever
	Rickettsia prowazekii	Humans, flying squirrels	Lice, fleas	Epidemic typhus
SAR11 cluster	*Pelagibacter ubique*	Free-living marine bacterium		None

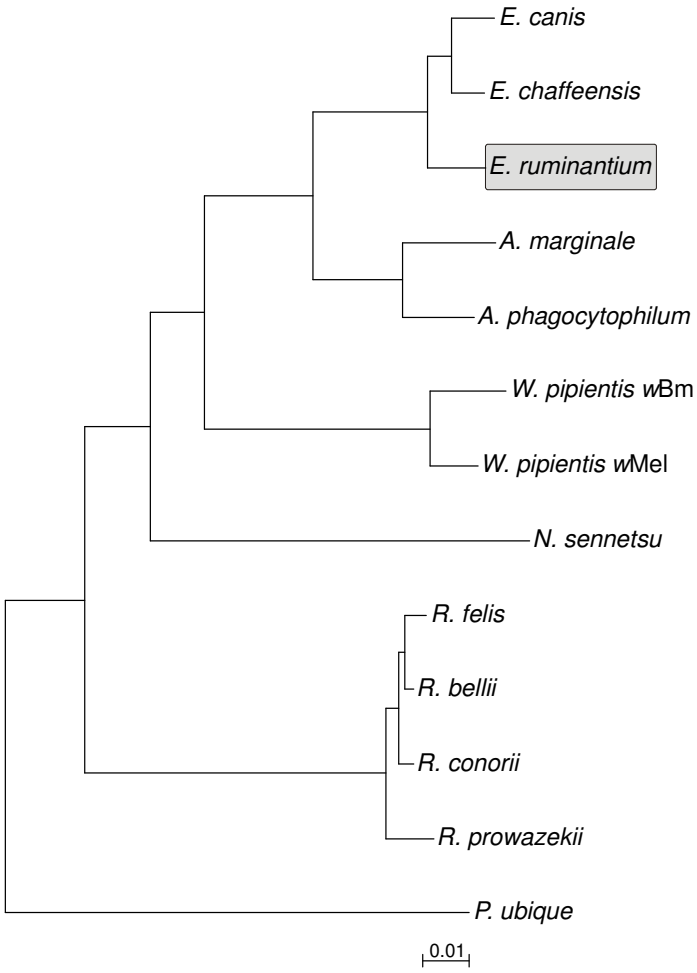


Figure 3.1. Neighbour-joining tree based on 16S rRNA sequences showing the phylogenetic relationships of *E. ruminantium* with other Rickettsiales for which complete genome sequences had been published at the time of this study. The sequences were aligned using ClustalX (Thompson *et al*., 2002) and the tree was inferred using the neighbour-joining method (Saitou & Nei, 1987).

3.2. MATERIALS AND METHODS

3.2.1. Metabolic reconstruction

Putative *E. ruminantium* metabolic pathways were analysed using the online pathway tools on the KEGG website (Ogata *et al.*, 1999; Kanehisa & Goto, 2000). All EC numbers were selected from the annotation and the list was then used to query the *E. coli* database. All the results were checked manually, and subsequently some gaps were filled by searching the similarity results in the annotation. We looked for ORFs with similar predicted products and functions for which no, or incorrect, EC numbers have been assigned. The pathways obtained from the KEGG website were reproduced in CorelDRAW® X3 (http://www.corel.co.uk).

3.2.2. *In silico* genome comparisons

The complete genome sequences of the organisms analysed in this study (Table 3.1) were retrieved from GenBank (ftp://ftp.ncbi.nih.gov/genbank/; accession numbers: *E. ruminantium* CR767821, *E. canis* CP000107, *E. chaffeensis* CP000236, *A. marginale* CP000030, *A. phagocytophilum* CP000235, *N. sennetsu* CP000237, *W. pipientis w*Mel AE017196, *W. pipientis w*Bm AE017321, *R. bellii* CP000087, *R. conorii* AE006914, *R. felis* CP000053, *R. prowazekii* AJ235269, *P. ubique* CP000084). Whole-chromosome alignments were done locally using Blastall (freely available at ftp://ftp.ncbi.nih.gov/blast) with default BLASTn parameters (Altschul *et al.*, 1990). The tabular view option (-m = 8) was used to allow visualisation of the alignments in the Artemis Comparison Tool (ACT) program (Carver *et al.*, 2005). The program formatdb, also included in the Blastall package, was used to convert Fasta files to BLAST databases.

All predicted *E. ruminantium* CDSs were translated and compared against the complete set of translated CDSs from each of the other 12 genomes. BLAST databases were created with formatdb from the predicted amino acid sequences of all CDSs, selected from GenBank files, of the 12 other Rickettsiales used for comparison. Unique and orthologous *E. ruminantium* genes

were identified by reciprocal BLASTp searches using parameters K = 10, b = 1 and an Expectation (E) value of 1. Similarity data were sorted with MSPcrunch (Sonnhammer & Durbin, 1994) using the default parameters. Homologous genes were identified as being the highest scoring hits which again yielded the original queries as the highest scoring hits in the reverse search direction. Only those pairs of homologous genes with a predicted amino acid identity ≥30% were retained for further analysis.

3.3. RESULTS AND DISCUSSION

3.3.1. Pathway analysis

3.3.1.1. Central metabolic pathways

3.3.1.1.1. Carbohydrate metabolism

Reconstruction of the central metabolic pathways (Figure 3.2) of *E. ruminantium* depicts an aerobic organism which probably does not ferment carbohydrates such as glucose, as many of the essential genes for the glycolytic pathway (e.g. hexokinase or glucokinase, and phosphofructokinase) were absent and a glucose transport system was not detected. An incomplete set of enzymes for glycolysis was also identified in the genomes of the other *Ehrlichia*, *Wolbachia* and *Anaplasma* species (Wu *et al*., 2004; Brayton *et al*., 2005; Foster *et al*., 2005; Hotopp *et al*., 2006; Mavromatis *et al*., 2006). We could not identify any enzymes for the Entner-Douderoff pathway, which is an alternative degradative pathway for carbohydrates in some microorganisms. The primary carbon sources are likely to be proline and glutamate, a prediction supported by the observation that the proline consumption of *E. ruminantium*-infected mammalian cells is increased in comparison with uninfected cells (Josemans & Zweygarth, 2002). Enzymes for the conversion of proline to glutamate were identified, including pyrroline-5-carboxylate reductase (*proC*) and the bifunctional enzyme proline dehydrogenase/delta-1-pyrroline-5-carboxylate dehydrogenase (*putA*). Probable transporters for both proline (*proP,* Erum 1330) and glutamate (sodium:dicarboxylate symporter family protein, Erum4510) were also identified.

Genes encoding all enzymes in the tricarboxylic acid (TCA) pathway were identified (Figure 3.3). A putative glutamate dehydrogenase (Erum3230) was identified that could feed glutamate into the TCA cycle through the reversible oxidative deamination of glutamate to α-ketoglutarate and ammonia. There was also a complete set of enzymes for the conversion of glutamate to fumarate and/or arginine. Enzymes for an intact pathway from pyruvate to fructose-6-phosphate were

identified (Figure 3.4); given the lack of a glycolytic pathway, the organism probably uses these enzymes solely for gluconeogenesis.

All enzymes for the non-oxidative branch of the pentose-phosphate pathway (Figure 3.4), which ultimately produces ribose 5-phosphate, were present. Ribose 5-phosphate and its derivatives are components of such important biomolecules as ATP, CoA, NAD^+, FAD, RNA and DNA.

3.3.1.1.2. Nucleoside biosynthesis

Complete biosynthetic pathways for the synthesis of purine and pyrimidine nucleosides were identified (Figure 3.5), as in all the other members of the Anaplasmataceae (Wu *et al*., 2004; Brayton *et al*., 2005; Foster *et al*., 2005; Hotopp *et al*., 2006; Mavromatis *et al*., 2006). This is unusual for other intracellular pathogens, for example organisms in the Rickettsiaceae family (Min *et al*., 2008), and *Chlamydia trachomatis* (Stephens *et al*., 1998), lack the ability to synthesise nucleosides.

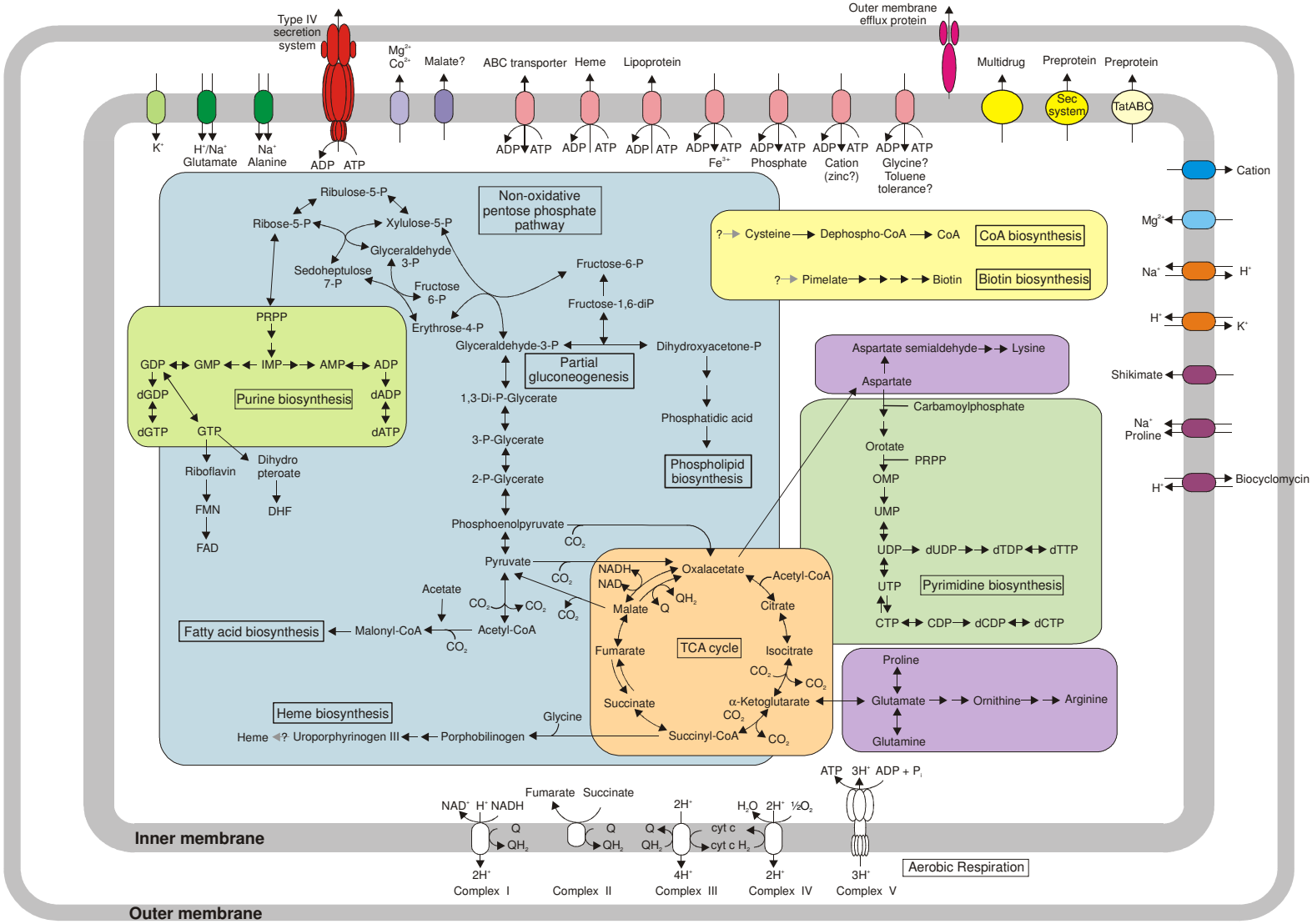


Figure 3.2. Schematic overview of metabolic pathways and substrate transport in *E. ruminantium.* Uncertainties are denoted by question marks. (Adapted from Collins *et al.,* 2005. [Supplementary information]).

3.3.1.1.3. Amino acid biosynthesis

The members of the Anaplasmataceae, particularly the *Ehrlichia* species, have a greater capacity to synthesise amino acids than *Rickettsia* species (Hotopp *et al*., 2006; Min *et al*., 2008). In *E. ruminantium* we identified genes encoding enzymes for the biosynthesis of the amino acids arginine, lysine, proline, glutamate and glutamine. Complete pathways for the biosynthesis of arginine from glutamate and lysine from aspartate could be established, as well as a pathway for interconversion between proline, glutamate and glutamine (Figure 3.3). The remaining 15 amino acids are likely to be obtained from the host cell, although we could only identify two specific transporters for proline (Erum1330) and glutamate (Erum4510). However, the components of several ATP-binding cassette (ABC) transporters were present, and it was not possible to identify the substrates for two of these. It is possible that these transporters have the ability to import a wide variety of substrates, which may include amino acids. In contrast, as expected for a free-living bacterium, *P. ubique* has complete biosynthetic pathways for all 20 amino acids (Giovannoni *et al*., 2005).

3.3.1.1.4. Cofactor biosynthesis

Several cofactor biosynthesis pathways were found (Figure 3.6), including those for biotin, coenzyme A and riboflavin. Genes encoding enzymes for dihydrofolate (DHF) synthesis were present, but we could not identify a gene coding for dihydrofolate reductase which is involved in the synthesis of tetrahydrofolate and folate from DHF.

All organisms in the Anaplasmataceae, with the exception of the *Wolbachia,* are able to synthesise cofactors and vitamins (Wu *et al*., 2004; Brayton *et al*., 2005; Foster *et al*., 2005; Hotopp *et al*., 2006; Mavromatis *et al*., 2006). Similarly to other endosymbionts, *W. pipientis* has completely lost the biosynthetic pathways for biotin, thiamine, and NAD (Foster *et al*., 2005). *R. prowazekii* has also lost the ability to synthesise biotin, thiamine, as well as NAD and, in addition, cannot synthesise FAD, pantothenate, and pyridoxine-phosphate (Andersson *et al*., 1998).

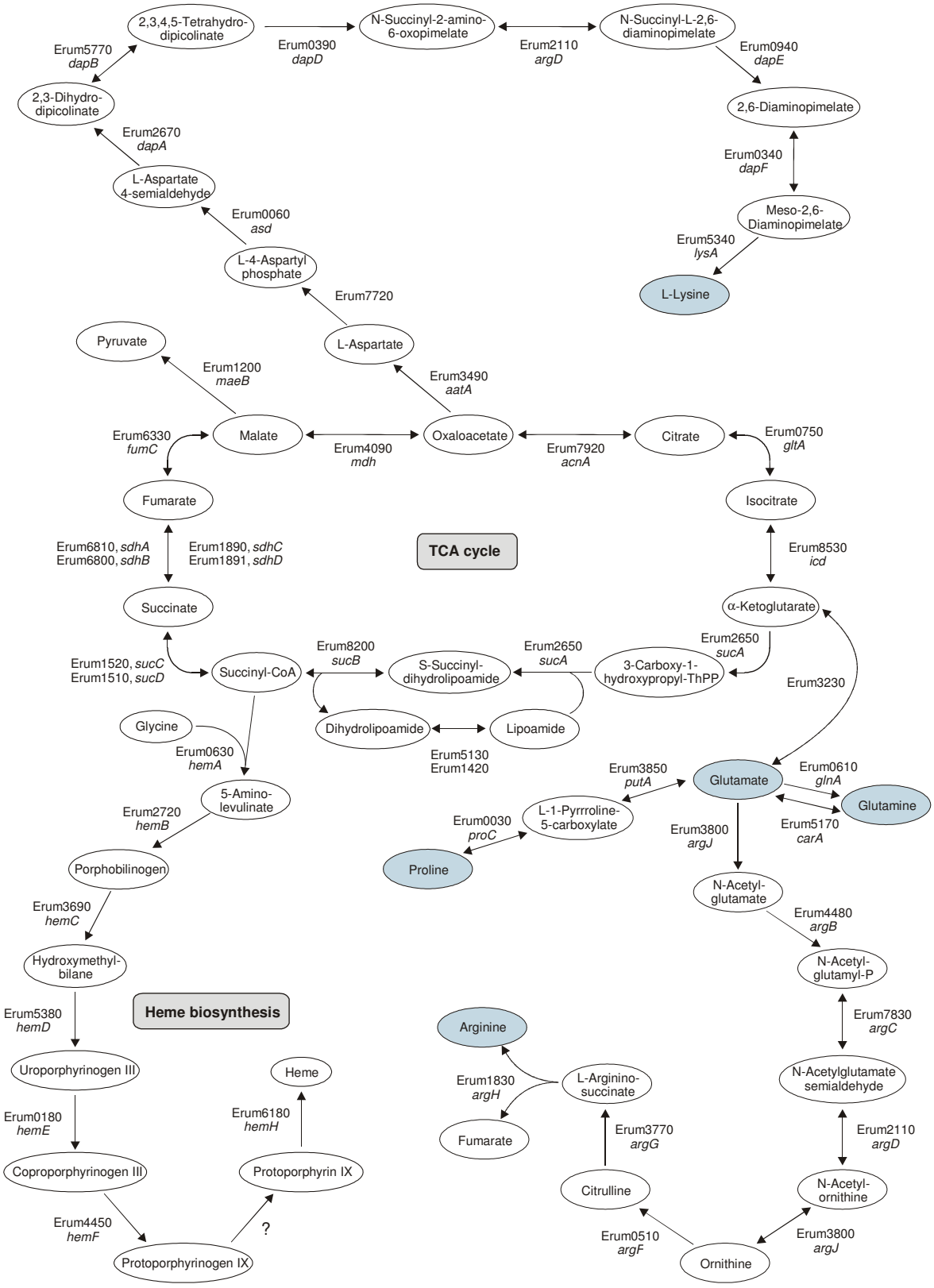


Figure 3.3. *E. ruminantium* genes coding for the enzymes involved in the TCA cycle, heme biosynthesis and amino acid biosynthesis. Amino acids for which pathways were identified are indicated in grey ovals.

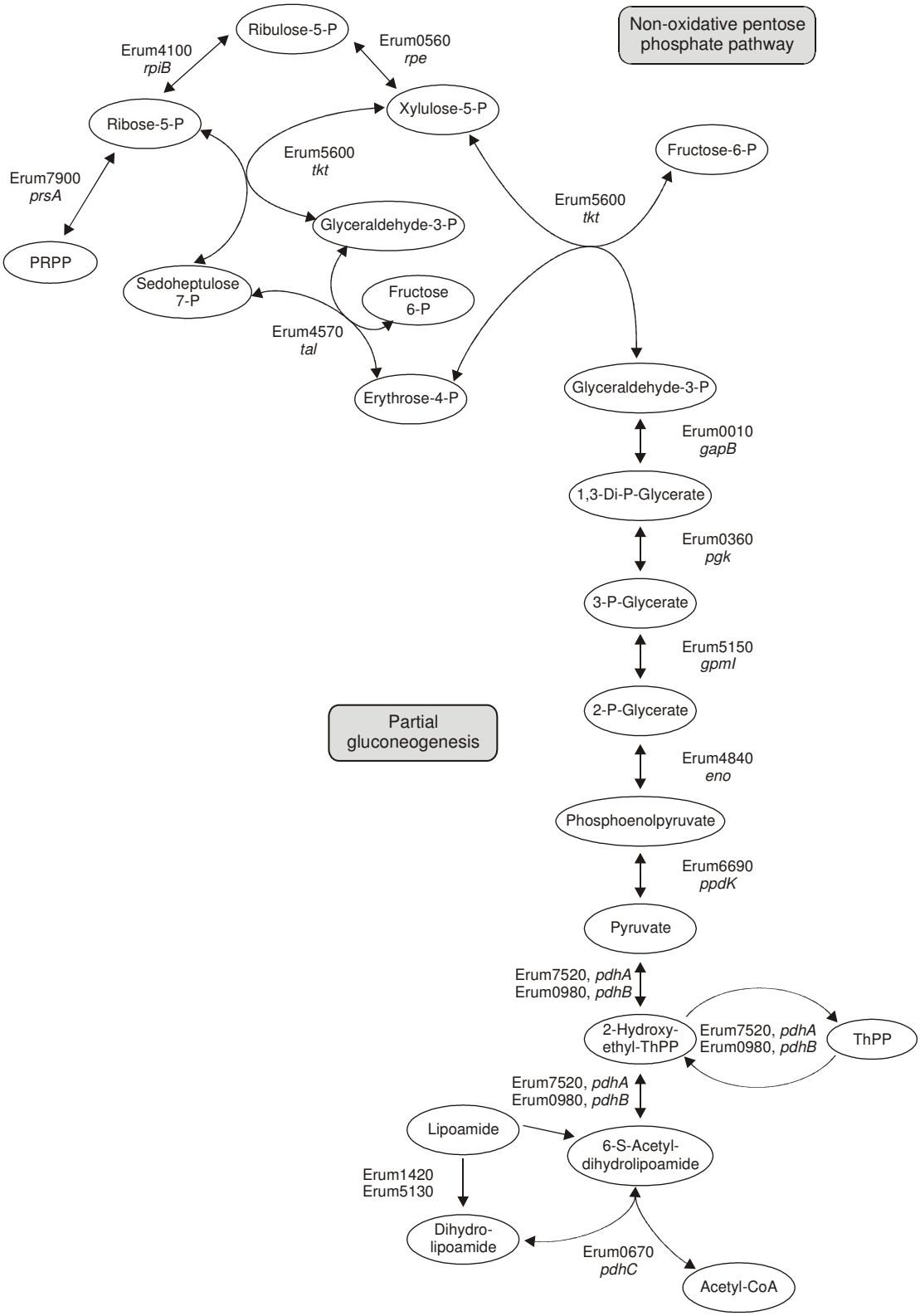


Figure 3.4. *E. ruminantium* genes involved in the pentose phosphate and gluconeogenesis pathways.

Hotopp and co-workers suggested that the presence of nucleotide, vitamin and cofactor biosynthetic pathways implies that *Anaplasma*, *Ehrlichia*, and *Neorickettsia* species do not compete with the host cell for, and may even supply host cells with, essential vitamins and nucleotides (Hotopp *et al*., 2006). Previously it has been proposed that the bacterial endosymbiont *Wigglesworthia glossinidia* supplies its host, *Glossina brevipalpis,* with as many as 60 vitamins that are rare in the blood meal of the tsetse fly (Zientz *et al*., 2004), and it is interesting that the cofactor and amino acid biosynthesis pathways of *Ehrlichia* and *Anaplasma* species are very similar to those of *W. glossinidia*. This is not to suggest, however, that *E. ruminantium* is a symbiote of *Amblyomma* ticks, the majority of which are not infected by the bacterium even in heartwater-endemic areas in Africa (Allsopp *et al*., 1999).

3.3.1.1.5. Lipid metabolism and cell wall components

Similarly to other members of the order Rickettsiales, *E. ruminantium* has genes for enzymes which perform fatty acid and phospholipid biosynthesis from intermediates of central metabolism, including those for phosphatidylglycerol and cardiolipin biosynthesis. No genes for enzymes essential for the production or modification of unsaturated fatty acids were identified.

No genes for lipopolysaccharide or peptidoglycan biosynthesis were identified in the *E. ruminantium* genome, and other members of the Anaplasmataceae family also lack these genes. The absence of such cell wall components, which impart strength and structure to the cell membranes of other Gram-negative bacteria, explains the fragile nature of the organism. *E. ruminantium* may use cholesterol from the host cell to compensate for the lack of lipid A and peptidoglycans, as has been shown to occur in *E. chaffeensis* and *A. phagocytophilum* (Lin & Rikihisha, 2003).

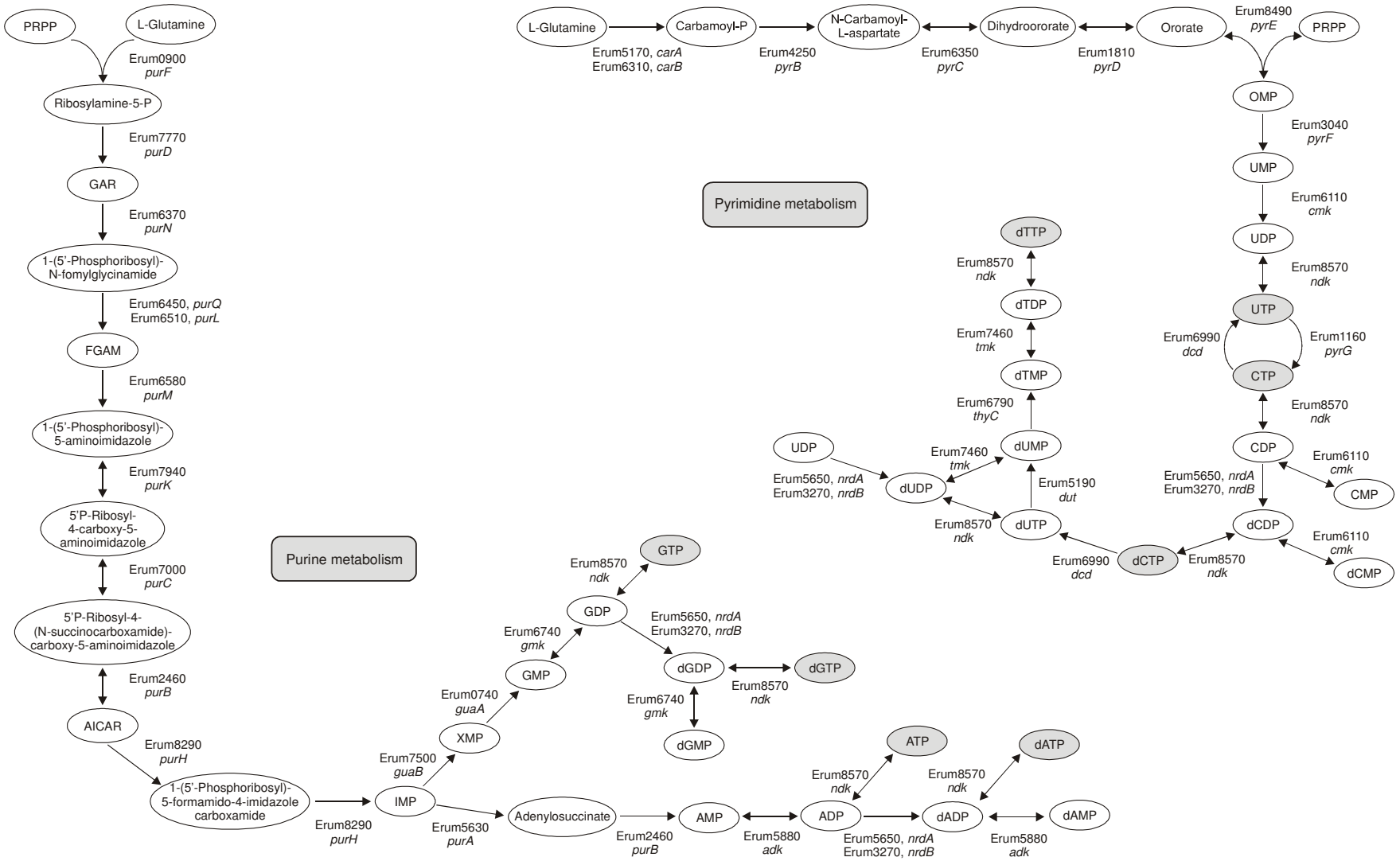


Figure 3.5. *E. ruminantium* genes involved in nucleotide metabolism.

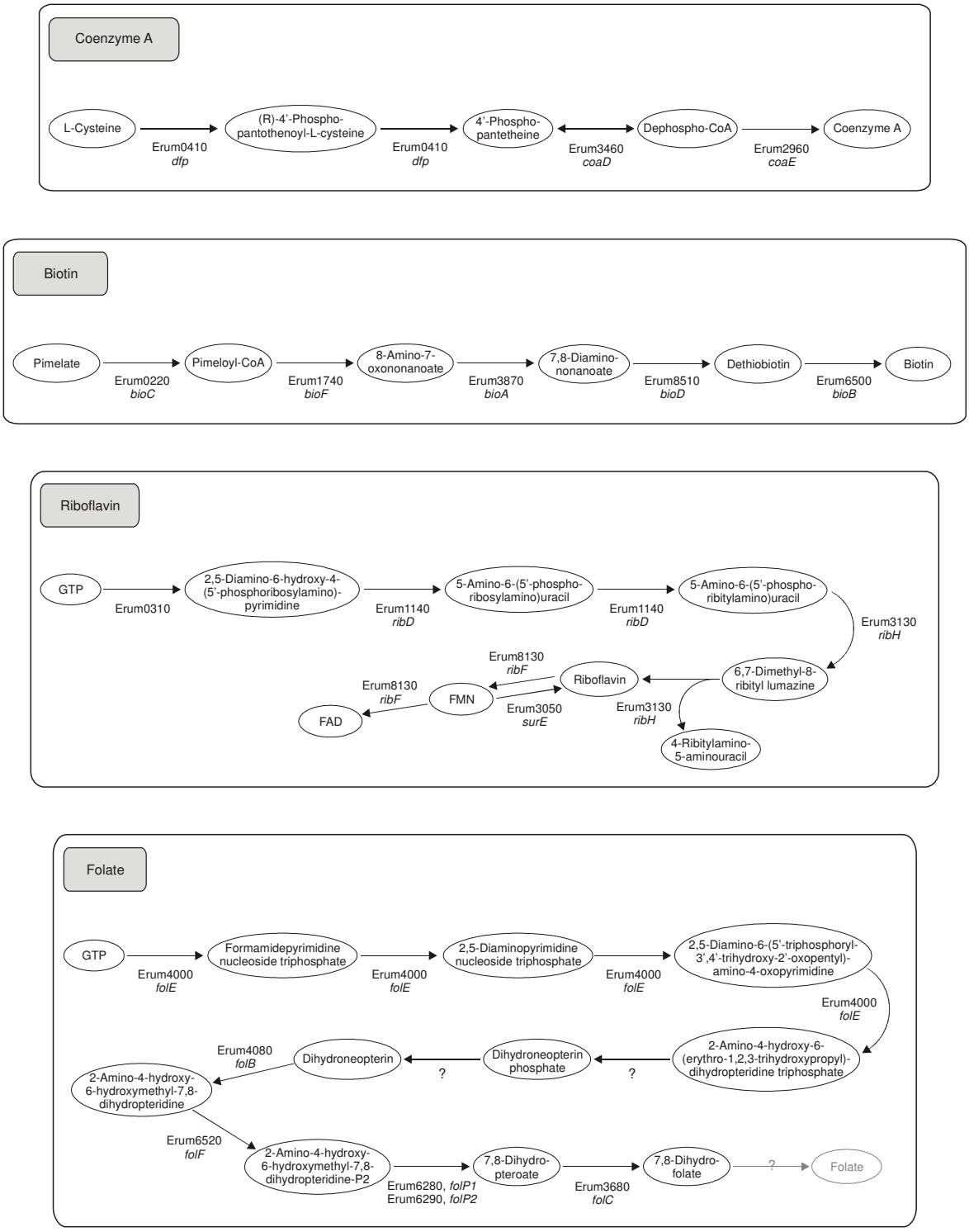


Figure 3.6. *E. ruminantium* genes involved in cofactor biosynthesis. Black question marks indicate enzymes that have not been characterised while uncertainties are denoted in grey.

3.3.1.2. Energy metabolism

E. ruminantium has several genes encoding putative enzyme complexes typical of aerobic respiration, including the ATP-synthase complex and the electron transfer complexes. The ATP-synthesizing complex produces ATP from ADP using energy from a proton gradient across the membrane. It is composed of two components: F_1, the catalytic core, and F_o, a hydrophobic segment that spans the membrane and forms the proton channel. The genes encoding these components are normally clustered in a single operon which is highly conserved in microbial genomes (Deckers-Hebestreit & Altendorf, 1996; Das & Ljungdahl, 2003). The ATP-synthase genes (*atpH, atpA, atpG, atpD* and *atpC*) encoding the α, β, γ, ε and δ subunits of the F_1 complex are located in three dispersed areas of the *E. ruminantium* genome in the following groups: (*atpH*, *atpA*), *atpG* and (*atpD*, *atpC*). The genes encoding the A, B and C chains of the F_o complex (*atpB, atpE, atpF*) are found clustered together.

ATP production is facilitated by a proton electrochemical gradient generated by an electron transport system consisting of NADH dehydrogenase (complex I), succinate dehydrogenase (complex II), cytochrome reductase (complex III) and cytochrome oxidase (complex IV). Genes coding for components of the NADH dehydrogenase complex are found in several clusters dispersed in the genome. Ten of these genes (*nuoGH*, *nuoDE* and *nuoNMLKJF*) are located near each other. There is an additional set of three genes grouped in the order *nuoABC*, with the single gene *nuoI* in between clusters *nuoABC* and *nuoGH*. Three additional individual genes closely related to *nuoM* are possible components of the NADH dehydrogenase complex. *E. ruminantium* succinate dehydrogenase consists of subunits similar to those found in *Campylobacter jejuni* (Parkhill *et al*., 2000), encoded by the genes *sdhA*, *sdhB*, *sdhC* and *sdhD*. Several proteins in the cytochrome bc_1 reductase complex, including ubiquinol-cytochrome *c* reductase iron-sulphur subunit (*petA*), cytochrome *b* (*petB*) and cytochrome c_1 (*petC*), were present, as were most subunits of the cytochrome oxidase complex (*coxA*, *coxB* and *coxC*). A complete pathway for porphyrin biosynthesis was identified, as well as several proteins responsible for cytochrome biosynthesis, supporting a central role for aerobic respiration and an electron transport system.

No ATP/ADP translocases were identified, which suggests that *E. ruminantium* does not make use of ATP from the host cell, unlike the related obligate intracellular parasites *R. prowazekii* (Andersson *et al.*, 1998) and *C. trachomatis* (Stephens *et al.*, 1998).

3.3.1.3. Replication, repair and recombination

As in the case of other intracellular organisms, *E. ruminantium* contains a small subset of the genes involved in DNA replication in free-living organisms (Andersson *et al.*, 1998; Akman *et al.*, 2002). Five genes which form the core structure of a functional DNA polymerase III were identified, these were *dnaE*, *dnaN*, *holB*, *dnaQ* and *dnaZ* putatively encoding the α, β, δ', ε and γ chains of the polymerase. There was also a gene encoding DNA polymerase I (*polA*). *E. ruminantium* DNA repair mechanisms appear to be similar to those found in other intracellular parasites, and several DNA repair genes were found, such as *mutM*, *radA*, *radC* and *nth* and the transcription-repair coupling factor *mfd*. Mismatch-repair enzymes were limited to *mutS* and *mutL*, and only one gene of the ultraviolet-induced DNA damage repair system (*uvrABC*), encoding subunit A, was identified. *E. ruminantium* has several genes involved in homologous recombination, such as *rmuC*, *recA*, *recR*, *recF* and a gene similar to *recO* (Erum4920) of *Mesorhizobium loti* (Kaneko *et al*., 2000). Although a gene coding for an enzyme similar to *recB* (Erum6250) was identified, the *recBCD* complex was missing.

3.3.1.4. Transcription and translation

We identified the DNA-dependent RNA polymerase of *E. ruminantium*, which consists of four subunits (α, β, β' and ω) encoded by *rpoA*, *rpoB*, *rpoC* and *rpoZ*. There were also two initiation factors σ^{70} and σ^{32} encoded by *rpoD* and *rpoH*. The *nusA*, *nusG*, *greA* and *rho* genes involved in transcription elongation and termination were also present. There were two very similar copies of the *rho* gene; *rho1* was 60 base pairs longer than *rho2* at the 5' end, where there were also several nucleotide differences. Several genes involved in RNA degradation were identified, including *rnpA* and *rnpB* (ribonuclease P), and *rnhA*, *rnhB* and *rnc*, encoding ribonucleases HI, HII and III respectively.

There is a single copy of each of the rRNA genes, which have a much higher G+C content than the rest of the genome (48.6%, 49.6% and 45.8% for 16S, 5S and 23S rRNA genes respectively). The 16S rRNA gene is widely separated from the 5S and 23S rRNA gene cluster. Several genes involved in rRNA processing and modification were found, including *ksgA*, *rbfA*, *rimM* and two pseudouridine synthetases, *rluC* and *rluD*. *E. ruminantium* contains a complete set of ribosomal proteins, except for the 50S ribosomal protein L30; in *E. coli* this protein is encoded by *rpmD* (Cerretti *et al.*, 1983) which we were not able to identify.

We identified 36 tRNA genes with specificities for all 20 amino acids, and several genes for tRNA modification were found, including *truB*, *miaA*, *rnpA* and *trmD*. Aminoacyl-transfer RNA (tRNA) synthetase genes were present for the aminoacylation of nearly all amino acids, including two genes encoding glutamyl-tRNA synthetase (*gltX1* and *gltX2*). Similarly to several other bacterial genomes, the genes encoding glutaminyl-tRNA synthetase and asparaginyl-tRNA synthetase were absent (Ibba *et al.*, 1997). Putative genes (*gatA*, *gatB* and *gatC*) coding for the three subunits of glutamyl-tRNA amidotransferase were identified, suggesting that the organism derives glutaminyl-$tRNA^{Gln}$ and asparaginyl-$tRNA^{Asn}$ by transamidation of mis-acylated glutamyl-$tRNA^{Gln}$ and aspartyl-$tRNA^{Asn}$. A putative tmRNA was found, responsible for tagging incomplete proteins on stalled ribosomes during proteolysis.

3.3.2. Transporters

The *E. ruminantium* genome sequence revealed numerous orthologs involved in eubacterial membrane transport systems (Figure 3.2). Several of these are ATP-binding cassette (ABC) transporters putatively involved in transportation of glycine, phosphate, lipoprotein, heme and ferric iron and other cations. Several different transporters involved in import and efflux of cations were identified. Na^+/H^+ (Erum1780, Erum5530 and Erum5550) and K^+/H^+ (Erum0950) antiporters are probably involved in maintaining the pH of the *E. ruminantium* cell. We found two transporters putatively involved in multidrug efflux, which may be responsible for the export

of anti-microbial host cell products. Our analyses indicate that *E. ruminantium* has the same basic mechanisms of secretion as those found in other free-living proteobacteria, these include common chaperones such as *dnaK*, *dnaJ*, *hslU*, *hslV*, *groEL*, *groES* and *htpG,* genes of the *secA*-dependent secretion system, and the *sec*-independent secretion system, *tat*.

3.3.3. Synteny analysis

Whole genome alignment can only be performed successfully for organisms that are sufficiently close phylogenetically, and we aligned *E. ruminantium* with the other twelve genome sequences to determine the degree of gene order conservation. Figures 3.7-11 represent the alignments displayed in ACT. The grey bars in the images represent the forward and reverse strands of DNA with the scale marked in base pairs. The coloured lines drawn between two adjacent linearised chromosomes show the location of homologous genes and indicate the same (red) or opposite (blue) orientation relative to the chromosome immediately above.

Large-scale gene order conservation across the chromosomes was found when the three *Ehrlichia* species were aligned (Figure 3.7), and a single symmetrical inversion near two duplicated genes which distinguishes *E. chaffeensis* from the other two *Ehrlichia* species will be discussed in Chapter 4. None of the other genera displayed the degree of synteny between species within each genus that was found within the *Ehrlichia* genus. Little conservation of gene order was found between *E. ruminantium* and the *Anaplasmas* (Figure 3.8), while there was no significant synteny between *E. ruminantium* and the *Wolbachia* (Figure 3.9) species, although these organisms have much in common with *E. ruminantium* as far as gene content is concerned. More than 75% of the predicted *E. ruminantium* ORFs have orthologs in the *Anaplasma* genomes, while 65-68% of the *E. ruminantium* genes share significant similarity with *Wolbachia* ORFs (Table 3.2). This observation correlates with the fact that *Anaplasma* species are phylogenetically closer to *E. ruminantium* than *Wolbachia* (Figure 3.1). No synteny was observed when we compared *E. ruminantium* with *N. sennetsu,* the *Rickettsia* species, and *P. ubique* (Figure 3.10, 3.11).

3.3.4. Shared and genus-specific genes

In total 33.6% of the *E. ruminantium* ORFs are conserved in all the genera we studied, including the free-living *P. ubique*, and a further 10.6% are found in all the Rickettsiales excluding *P. ubique* (Table 3.2). The conserved genes are generally associated with house keeping functions. Of the 888 predicted protein coding sequences in *E. ruminantium* 99 (11.1%) are unique to this species. The products of these genes are unknown, but 60 are predicted to be membrane-associated, six are probably exported, and some are likely to be involved in niche adaptation and pathogenic characteristics. Seven percent of the *E. ruminantium* ORFs, all of unknown function, are shared only with other *Ehrlichia* species, 42 ORFs (4.7%) are shared by *Ehrlichia* and *Anaplasma* species, while 11 genes are conserved between the genera *Ehrlichia* and *Wolbachia*. Five genes (*argC*, *argG*, *argH*, *argJ*, and *lysA*) involved in arginine and lysine biosynthesis are shared only by the *Ehrlichia* species and *P. ubique*, and Erum3980, an ORF containing ankyrin repeats, is found only in the *Ehrlichia* species and *N. sennetsu*.

Interestingly, some *E. ruminantium* ORFs are similar to predicted genes in one of the other genera, but are not shared with *E. chaffeensis* or *E. canis* (Table 3.2). For example, three ORFs (Erum0060, Erum2300 and Erum2410) have orthologs in only one of the *Rickettsia* species, and five (Erum1050, Erum2810, Erum4210, Erum7990 and Erum8000) are shared only by an *Anaplasma* species. Most of these are predicted to encode membrane proteins of unknown function except for Erum2810, which is a sugar transport protein.

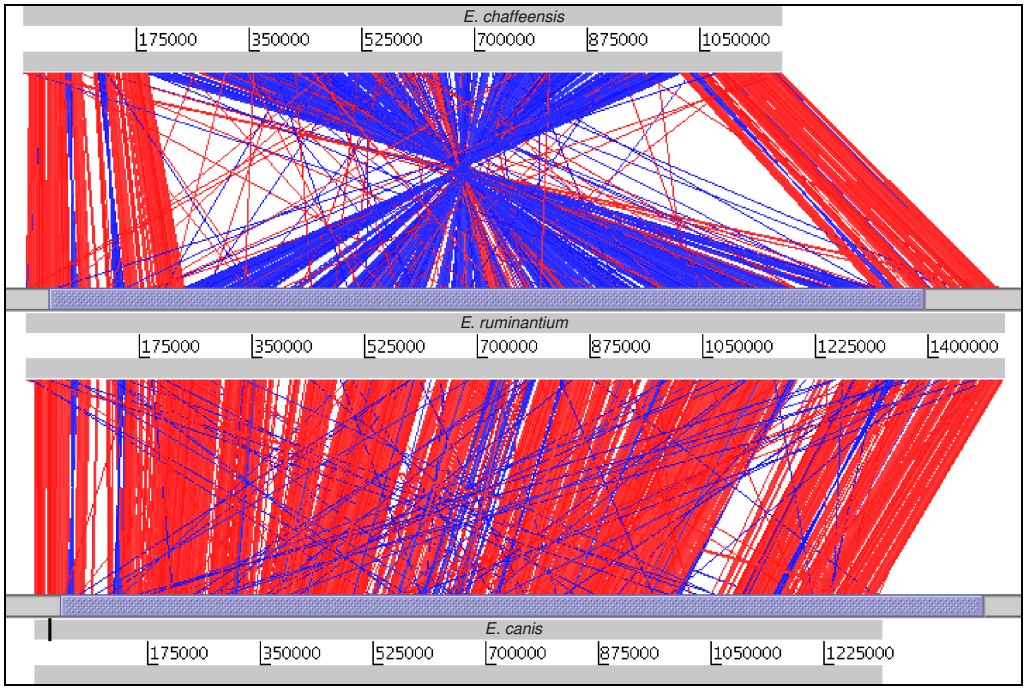


Figure 3.7. Global comparison between *E. ruminantium* (middle), *E. chaffeensis* (top) and *E. canis* (bottom) displayed using ACT.

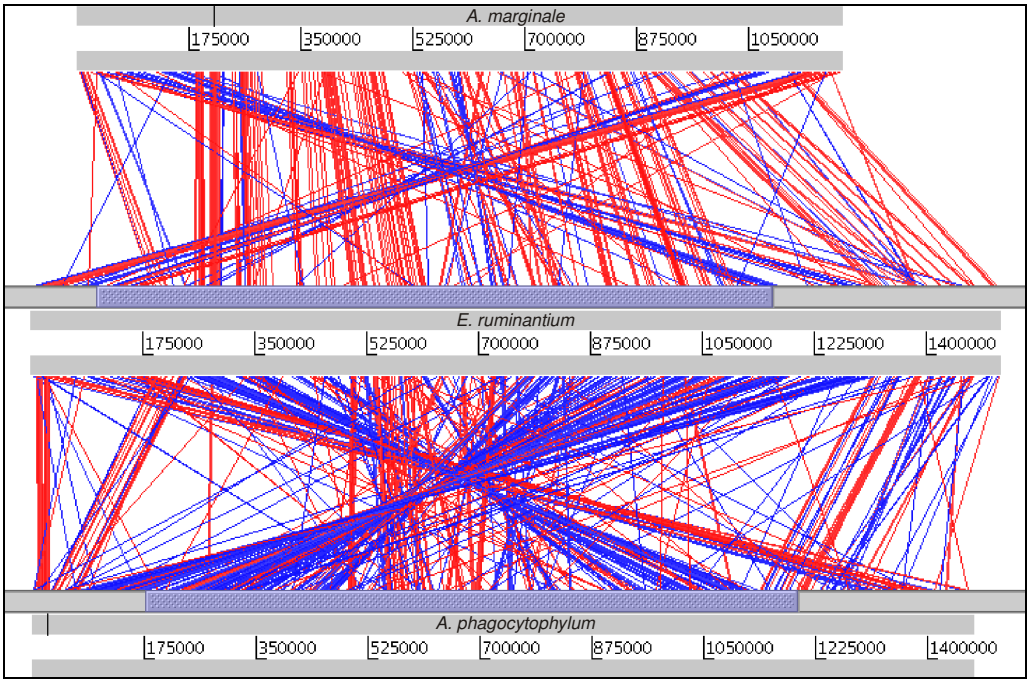


Figure 3.8. Comparison of chromosomal synteny between *E. ruminantium* (middle), *A. marginale* (top) and *A. phagocytophilum* (bottom).

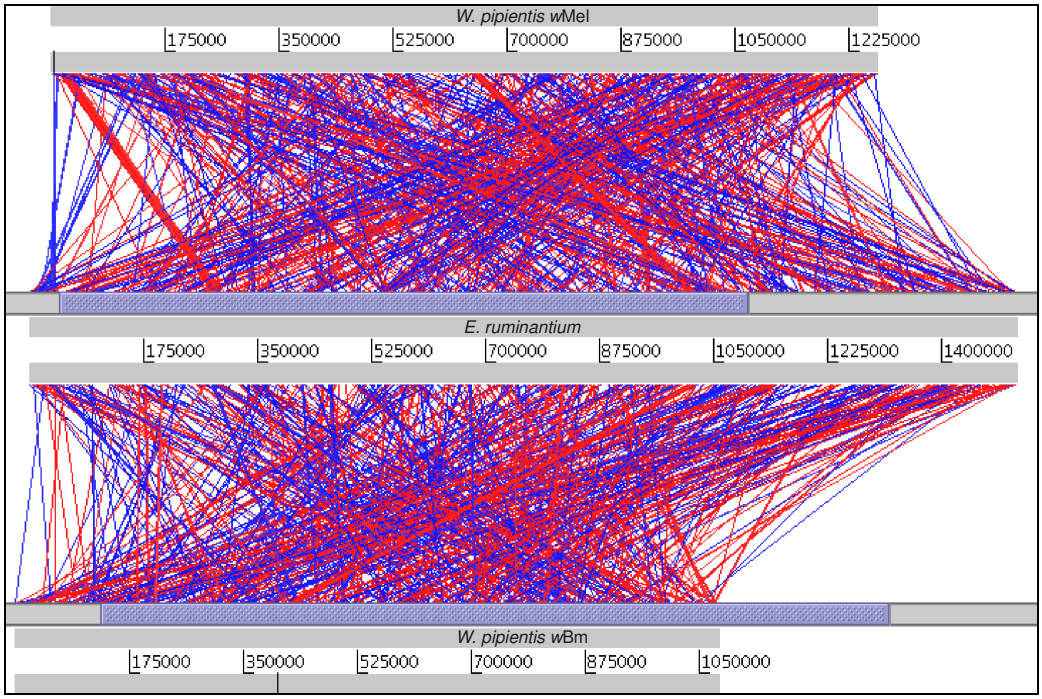


Figure 3.9. Genomic location of the homologous genes in *E. ruminantium* (middle) and the two *Wolbachia* species.

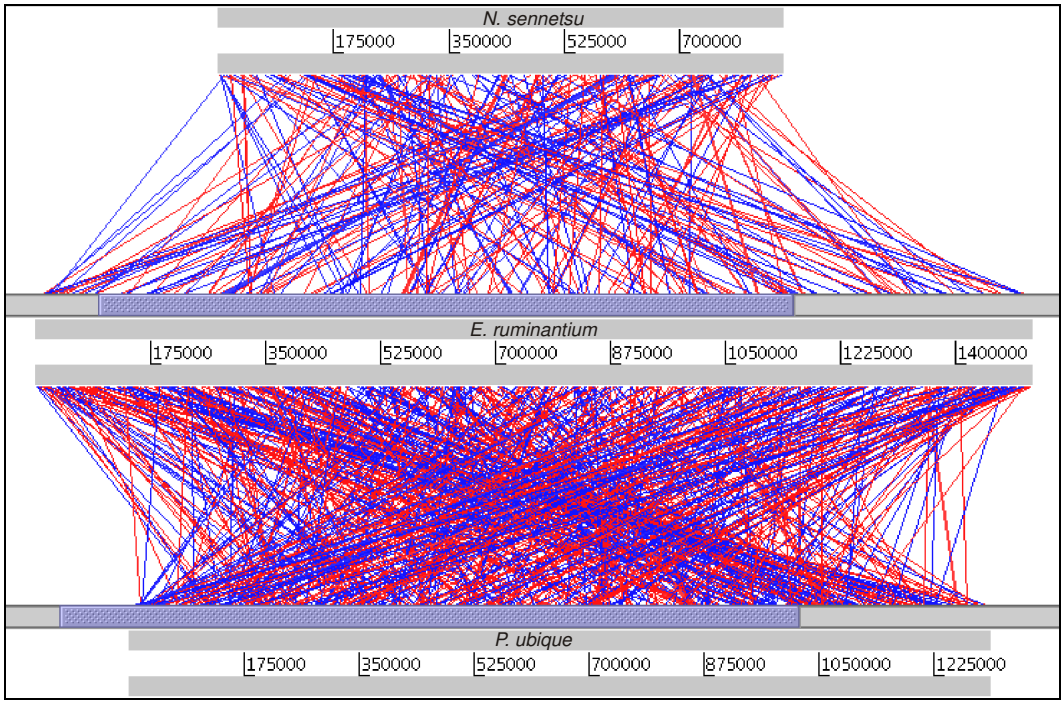


Figure 3.10. *E. ruminantium* gene order compared to *N. sennetsu* (top) and *P. ubique* (bottom).

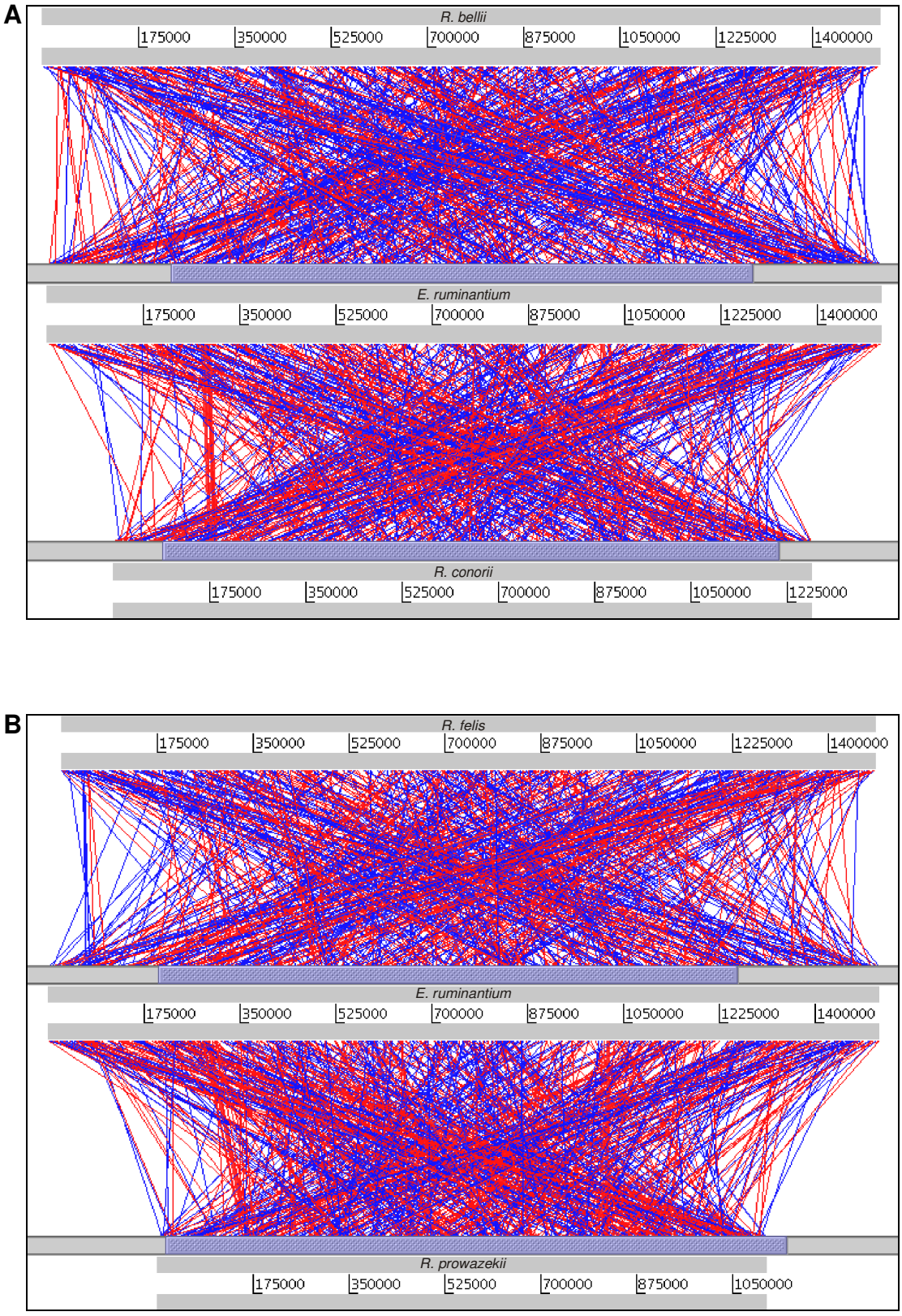


Figure 3.11. **A.** Comparison of relative positions of conserved genes between *E. ruminantium*, *R. bellii* (top) and *R. conorii* (bottom). **B.** *E. ruminantium* gene order compared to *R. felis* (top) and *R. prowazekii* (bottom).

3.4. CONCLUSIONS

The genome-based metabolic reconstruction of *E. ruminantium* revealed the metabolic and biosynthetic capabilities typical of an organism having an obligate intracellular lifestyle. The ever-increasing number of genome sequences of pathogens has provided us with an opportunity to use comparative genomic analysis to explore many of the aspects of the biology of the order Rickettsiales. We identified a number of genes unique to *E. ruminantium* and also genes shared with other members in the Rickettsiales. The challenge now is to reconcile the genomic differences and similarities with the observed variations in the vectors, host relationships and lifestyles of the different species. Since most of the genes that are not shared are not functionally characterised in any organism, further progress will only be made when this has been achieved. The ongoing accumulation of genomic data will certainly yield some of the required information, but it is also likely that specific *in vitro* expression characterisation experiments will have to be conducted for many of these unknown genes.

Table 3.2. *E. ruminantium* genes shared by other Rickettsiales. The first column represents the systematic identification number of *E. ruminantium* ORFs. Plus signs in columns 2-13 indicate the presence of *E. ruminantium* homologs in other species: Eca = *E. canis*, Ech = *E. chaffeensis*, Ama = *A. marginale*, Aph = *A. phagocytophilum*, WBm = *W. pipientis w*Bm, WMel = *W. pipientis w*Mel, Nsen = *N. sennetsu*, Rbel = *R. bellii*, Rcon = *R. conorii*, Rfel = *R. felis*, Rpro = *R. prowazekii*, Pub = *P. ubique*. See Appendix E for the annotation of each ORF.

Erum	Eca	Ech	Ama	Aph	WBm	WMel	Nsen	Rbel	Rcon	Rfel	Rpro	Pub
0010	+	+	+	+	+	+	+					+
0020	+	+	+	+	+	+						
0030	+	+										
0040	+	+	+	+	+	+	+	+	+	+	+	+
0050	+	+	+	+	+	+	+	+	+	+	+	+
0060	+	+	+	+	+	+	+	+	+	+	+	+
0070	+	+	+	+	+	+	+			+		+
0080	+	+	+	+	+	+	+					
0090	+	+			+							
0110	+	+	+	+	+	+	+		+	+	+	+
0120	+	+	+	+	+	+	+	+	+	+	+	
0130	+	+	+	+	+	+	+	+	+	+	+	+
0140	+	+	+	+			+					+
0150	+	+	+	+	+	+						+
0160	+	+	+	+	+	+	+	+	+	+	+	
0170	+	+	+	+	+	+	+	+	+	+	+	+
0180	+	+	+	+	+	+	+	+	+	+	+	+
0190	+	+	+	+	+	+	+	+	+	+	+	
0200	+	+	+	+		+	+			+		
0210	+	+	+	+		+	+	+	+	+		
0220	+	+	+	+								
0230	+	+	+	+			+					+
0240	+	+	+	+	+	+	+	+	+	+	+	+
0260	+	+	+	+	+	+	+	+	+	+	+	
0270	+	+	+	+	+	+	+	+	+	+	+	+
0280	+	+	+	+	+	+	+	+	+	+	+	
0290	+	+	+	+	+	+	+		+	+	+	
0300	+	+	+	+	+	+	+		+		+	
0310	+	+	+	+	+	+	+					
0320	+	+	+	+	+	+	+	+				
0330	+	+										
0340	+	+	+		+	+		+	+	+	+	+
0350	+	+	+	+		+						
0360	+	+	+	+	+	+	+					+
0370	+	+	+	+				+	+	+	+	
0380	+	+	+	+		+		+	+	+	+	+
0390	+	+	+		+	+		+	+	+	+	+
0400	+	+	+	+	+	+	+	+	+	+	+	+
0410	+	+	+	+								
0420	+	+	+	+	+	+	+	+	+	+	+	+
0430	+	+	+	+	+	+	+	+		+		+
0440	+	+	+	+	+	+	+	+	+	+	+	+
0450	+	+	+	+	+							
0460	+	+	+	+	+	+						
0470	+	+	+	+								
0480	+	+	+	+	+	+	+	+	+	+	+	+
0490	+	+	+	+	+	+	+	+	+	+	+	+
0510	+	+	+									+
0520	+	+	+	+	+	+		+	+	+	+	+
0530	+	+	+	+	+	+	+	+	+	+	+	
0540	+	+	+	+	+	+	+	+	+	+	+	+
0550	+	+	+	+	+	+	+	+	+	+	+	+
0560	+	+	+	+	+	+	+					+
0570	+	+	+	+	+	+	+	+	+	+	+	
0580	+	+	+	+	+	+	+	+	+	+	+	
0590	+	+										
0600	+	+	+	+	+	+	+	+	+	+	+	+
0610	+	+	+	+			+					+
0620	+	+	+	+	+	+	+	+	+	+	+	+
0630	+	+	+	+	+	+	+	+	+	+	+	
0631	+	+	+		+		+					

Erum	Eca	Ech	Ama	Aph	WBm	WMel	Nsen	Rbel	Rcon	Rfel	Rpro	Pub
0640	+	+	+	+	+	+	+	+	+	+	+	+
0650	+	+	+	+	+	+	+					
0660									+			
0670	+	+	+	+	+	+	+	+	+	+	+	
0730	+	+	+	+								
0740	+	+	+	+	+	+	+					+
0750	+	+	+	+	+	+	+	+	+	+	+	+
0770	+	+	+	+	+	+	+					
0780	+	+	+	+	+	+	+	+	+	+	+	
0790	+	+	+	+	+	+	+	+	+	+	+	+
0800	+	+	+	+	+	+	+					+
0810	+	+	+	+	+	+	+	+	+	+	+	+
0820	+	+	+	+	+	+	+	+	+	+	+	+
0830	+	+			+	+	+					
0831	+	+	+	+								
0840	+	+		+		+						
0850	+	+										
0860	+	+	+	+	+	+	+	+	+	+	+	+
0870	+	+	+	+	+	+						
0880	+	+	+	+	+	+	+	+	+	+	+	+
0890	+	+	+	+	+	+						
0900	+	+	+	+	+	+	+					+
0910	+	+	+	+	+	+	+	+	+	+	+	+
0920	+	+	+	+	+	+	+	+	+	+	+	+
0930	+	+	+	+		+	+	+	+	+		
0940	+	+	+		+	+		+	+	+	+	+
0950	+	+	+	+	+	+	+	+	+	+	+	+
0960	+	+	+	+	+	+	+					+
0970	+	+	+	+	+	+	+	+	+	+	+	
0980	+	+	+	+	+	+	+	+	+	+	+	
1000	+	+	+	+	+	+	+					+
1010	+	+	+	+	+	+	+	+	+	+	+	+
1020	+	+	+	+	+	+	+					+
1030	+	+	+	+			+					
1050				+								
1060	+	+	+	+	+	+	+					+
1070	+	+	+	+								
1080	+	+	+	+	+	+			+	+		+
1090	+	+	+	+	+	+	+		+	+	+	+
1120	+	+	+	+	+	+	+	+	+	+	+	+
1130	+	+	+	+	+	+	+	+	+	+	+	+
1140	+	+	+	+	+	+	+					+
1160	+	+	+	+	+	+	+	+	+	+	+	+
1170	+	+	+	+	+	+						+
1180	+	+	+	+	+	+	+	+	+	+	+	
1190	+	+	+	+	+	+	+	+	+	+	+	+
1200	+	+	+	+	+	+	+	+	+	+	+	+
1210	+	+	+									
1220	+	+					+	+			+	
1240	+	+	+	+	+	+	+					
1250	+											
1260	+	+	+	+	+	+	+	+	+	+	+	
1270	+	+		+	+	+	+	+	+	+	+	+
1280	+	+	+	+			+					+
1290	+	+										
1300	+	+	+	+		+						
1310	+	+	+	+	+	+	+					+
1320	+	+	+	+	+	+	+	+	+	+	+	+
1330	+	+	+	+	+	+	+	+	+	+	+	
1340	+	+	+	+								
1350	+	+	+	+			+					+
1360	+	+	+	+	+	+	+	+	+	+	+	+
1370	+	+	+	+	+	+	+	+	+	+	+	+
1380	+	+	+	+	+	+	+	+	+	+	+	+
1390	+	+	+	+	+	+	+	+	+	+	+	+
1400	+	+	+	+	+	+	+	+	+	+	+	+
1420	+	+	+	+	+	+	+	+	+	+	+	
1430	+	+										
1440	+	+	+									
1450	+	+	+	+	+	+						
1460	+	+	+									
1470	+	+	+	+			+	+	+	+	+	+
1480	+	+	+	+	+	+	+	+	+	+	+	
1490	+	+	+	+	+	+	+					+
1500	+	+	+	+	+	+	+	+	+	+	+	+

Erum	Eca	Ech	Ama	Aph	WBm	WMel	Nsen	Rbel	Rcon	Rfel	Rpro	Pub
1510	+	+	+	+	+	+	+	+	+	+	+	+
1520	+	+	+	+	+	+	+	+	+	+	+	+
1530	+	+	+	+	+	+	+					+
1540	+	+	+	+	+	+	+					+
1550	+	+	+	+	+	+	+	+	+	+	+	
1560	+	+	+	+	+	+	+					
1570	+	+	+	+	+	+	+					
1580	+	+	+	+	+	+	+					+
1590	+	+			+	+	+	+	+	+	+	
1600	+	+	+	+	+	+						
1610	+	+	+	+	+	+	+	+	+	+	+	+
1620	+	+										
1630	+	+	+	+	+	+	+	+	+	+	+	+
1640	+	+	+	+	+	+	+	+	+	+	+	+
1650	+	+	+	+	+	+	+	+	+	+	+	+
1660	+	+	+	+	+	+						
1670	+	+	+	+	+	+		+	+	+	+	+
1680	+	+	+	+	+	+	+	+	+	+	+	+
1690	+	+	+	+	+	+	+	+	+	+	+	+
1700	+	+	+	+	+	+						+
1710	+	+	+	+	+	+					+	+
1720	+	+	+	+			+	+	+	+	+	+
1730	+	+	+	+	+	+	+	+	+	+	+	+
1740	+	+	+	+								
1750	+	+	+	+	+	+		+	+	+	+	
1760	+	+	+	+	+	+	+	+	+	+	+	+
1770	+	+			+	+						
1780	+	+	+	+				+	+	+	+	
1790	+	+										
1800	+	+	+	+								+
1810	+	+	+	+	+	+	+					+
1820	+	+	+	+								
1830	+	+										+
1840	+		+			+						
1850	+	+	+	+	+	+	+					+
1851	+	+	+	+	+	+	+	+	+	+	+	
1860	+	+	+									
1870	+	+	+	+	+	+	+	+	+	+	+	+
1880	+	+		+								
1890	+	+	+	+	+	+	+	+	+	+	+	+
1891	+	+	+		+	+		+				
1900	+	+										
1910	+	+	+	+			+					+
1920	+	+	+	+	+	+		+				
1930	+	+	+	+	+	+						
1940	+	+	+	+	+	+	+	+	+	+	+	+
1950	+	+	+	+	+	+	+	+				
1960	+	+	+	+	+	+						
1970	+	+	+	+	+	+	+			+		
1980	+	+		+		+	+					
1990	+	+	+	+	+	+						
2000	+	+	+	+	+	+	+	+	+	+	+	+
2010	+	+	+	+	+	+	+	+	+	+	+	+
2020	+	+	+	+	+	+	+	+	+	+	+	+
2030	+	+	+	+	+	+	+	+	+	+	+	+
2040	+	+	+	+	+	+	+	+	+	+	+	+
2050	+	+	+	+	+	+	+	+	+	+	+	+
2060	+	+	+	+			+					+
2070	+	+	+	+	+	+	+	+	+	+	+	
2080	+	+	+	+	+	+		+				+
2090	+	+	+	+	+	+	+	+	+	+	+	+
2100	+	+										
2110	+	+	+	+	+	+	+					+
2120	+	+	+	+	+	+	+	+	+	+	+	
2130	+	+	+	+	+	+	+	+	+	+	+	
2140	+	+	+	+		+	+	+	+	+		
2150	+	+	+	+	+	+	+	+	+	+	+	+
2160	+	+	+	+	+	+	+	+	+	+	+	+
2180	+	+										
2190	+	+	+	+	+	+	+	+	+	+	+	+
2200	+	+	+	+	+	+						
2210	+	+	+	+	+	+	+	+	+	+	+	+
2220	+	+	+	+	+	+	+	+	+	+	+	+
2230	+	+	+	+	+	+	+	+	+	+	+	+
2280	+	+										

Erum	Eca	Ech	Ama	Aph	WBm	WMel	Nsen	Rbel	Rcon	Rfel	Rpro	Pub
2300								+				
2380	+	+										
2390	+	+	+	+	+	+	+	+	+	+	+	+
2410											+	
2420	+	+	+	+	+	+	+	+	+	+	+	+
2430	+	+	+	+	+	+	+	+	+	+	+	+
2440	+	+	+	+	+	+						
2450	+	+	+	+	+	+	+	+	+	+	+	
2460	+	+	+	+	+	+	+					+
2490		+										
2520	+	+	+	+	+	+	+	+	+	+	+	+
2530	+	+	+		+	+		+	+	+	+	
2540	+	+	+	+	+							
2550	+	+	+	+	+	+	+	+	+	+	+	+
2560	+	+		+	+	+	+	+	+	+	+	
2570	+	+	+	+	+	+	+	+	+	+	+	+
2580	+	+	+	+	+	+						
2590	+	+	+	+	+	+		+	+	+	+	
2600	+	+	+	+	+	+	+	+	+	+	+	+
2610	+	+				+						
2620	+	+	+	+	+	+	+	+	+	+	+	+
2630	+	+	+	+		+		+		+		
2640	+	+	+	+		+			+	+		
2650	+	+	+	+	+	+	+	+	+	+	+	+
2660	+	+	+	+		+	+	+		+		
2670	+	+	+				+	+	+	+	+	+
2680	+	+	+	+	+	+	+	+	+	+	+	+
2690	+	+										
2700	+	+	+	+	+	+	+	+	+	+	+	
2710	+	+	+	+			+					
2720	+	+	+	+	+	+	+	+	+	+	+	+
2730	+	+			+	+						
2740	+	+	+									
2810			+	+								
2820	+	+										
2830	+	+	+	+	+	+	+	+	+	+	+	+
2840	+	+	+	+	+	+						
2850	+	+	+	+	+	+	+	+	+	+	+	+
2860	+	+	+	+	+	+	+	+	+	+	+	+
2870	+	+	+	+	+	+	+	+	+	+	+	+
2900	+	+										
2910	+	+	+	+			+					+
2920	+	+	+	+		+	+					+
2930	+	+		+	+	+						
2940	+	+	+	+								
2950	+	+	+	+	+	+	+	+	+	+	+	+
2960	+	+	+	+	+	+						
2970	+	+	+	+								
2980	+	+	+	+	+	+	+					+
2990	+	+	+	+	+	+	+	+	+	+	+	+
3000	+	+	+	+	+	+	+					
3010	+	+	+	+	+	+	+	+	+	+	+	+
3030	+	+	+	+	+	+	+	+	+	+	+	+
3040	+	+	+	+	+	+	+					+
3050	+	+	+	+	+	+	+					
3060	+	+	+	+	+	+		+	+	+	+	+
3070	+	+	+	+	+	+	+	+	+	+	+	+
3090	+	+	+	+	+	+	+	+	+	+	+	+
3100	+	+	+	+	+	+	+	+	+	+	+	+
3110	+	+	+	+	+	+		+	+	+	+	+
3120	+	+	+	+	+	+	+					
3130	+	+	+	+	+	+	+					+
3140	+	+	+	+	+	+	+	+	+	+	+	+
3150	+	+			+	+						
3160	+	+	+	+	+	+	+	+	+	+	+	+
3170	+	+	+	+	+	+	+	+	+	+	+	+
3180	+	+	+	+	+	+	+	+	+	+	+	
3190	+	+	+	+	+	+	+	+	+	+	+	+
3200	+	+	+	+	+	+						+
3210	+	+	+	+	+	+	+	+	+	+	+	+
3220	+	+	+	+	+	+	+	+	+	+	+	
3221	+	+	+	+								
3230	+	+	+	+	+	+	+	+	+	+	+	
3240	+	+	+	+	+	+	+	+	+	+	+	
3250	+	+	+	+	+	+	+	+	+	+	+	+
3270	+	+	+	+	+	+	+					+

Erum	Eca	Ech	Ama	Aph	WBm	WMel	Nsen	Rbel	Rcon	Rfel	Rpro	Pub
3280	+	+	+	+	+	+	+	+	+	+	+	
3290	+	+										
3300	+	+	+	+	+	+	+	+	+	+	+	+
3310	+	+	+	+	+	+	+	+	+	+	+	
3320	+	+	+	+	+	+	+	+	+	+	+	+
3330	+	+	+	+	+	+	+	+	+	+	+	+
3340	+	+	+	+	+	+						
3350	+	+	+	+	+	+		+	+	+	+	
3360	+	+	+	+	+	+						
3370	+	+	+	+	+	+	+					+
3380	+	+	+	+	+	+						
3390	+	+	+	+								
3400	+	+	+	+	+	+	+	+	+	+	+	+
3410	+	+										
3420	+	+	+	+	+	+	+	+	+	+	+	
3430	+	+	+	+	+	+	+	+	+	+	+	+
3440	+	+	+	+	+	+	+	+	+	+	+	+
3450	+	+	+	+	+	+	+					
3460	+	+	+	+	+	+	+					+
3470	+	+	+	+	+	+	+	+		+	+	+
3480	+	+	+	+	+	+	+	+	+	+	+	
3490	+	+	+	+	+	+	+	+	+	+	+	+
3500	+	+	+	+	+	+						
3510	+	+	+	+	+	+		+	+	+	+	
3520	+	+	+	+		+	+	+	+	+	+	
3530	+	+	+	+	+	+	+	+	+	+	+	+
3540	+	+	+	+	+	+	+	+	+	+	+	+
3550	+	+	+	+								
3560	+	+	+	+	+	+	+	+	+	+	+	+
3640	+	+	+	+	+	+		+				+
3650	+		+	+	+						+	+
3660	+	+	+	+	+	+	+	+	+	+	+	+
3670	+	+	+	+	+	+	+	+	+	+	+	+
3680	+	+	+	+		+	+	+	+	+	+	+
3690	+	+	+	+	+	+	+	+	+	+	+	+
3700	+	+	+	+	+	+	+	+	+	+	+	+
3701	+	+										
3710	+	+	+	+	+	+	+	+	+	+	+	+
3720	+	+	+	+	+	+	+	+	+	+	+	+
3730	+	+	+	+	+	+	+		+	+	+	+
3740	+	+			+	+	+	+	+	+	+	+
3750	+	+					+				+	
3760	+	+										
3770	+	+										+
3780	+	+										
3790	+	+	+									
3800	+	+										+
3810	+	+	+	+	+	+	+	+	+	+	+	+
3820	+	+										
3830	+	+										
3840	+	+	+	+	+	+	+	+	+	+	+	+
3850	+	+	+	+	+	+						
3870	+	+	+	+			+					
3880	+	+	+	+		+	+	+	+	+		
3890	+	+		+			+					
3900	+		+									
3910	+	+				+						
3920	+	+				+						
3930	+	+	+	+	+	+						
3940	+	+		+		+						
3950	+	+	+	+	+	+	+	+	+	+	+	+
3960	+	+	+	+	+	+	+	+	+	+	+	+
3970	+	+	+									
3980	+	+					+					
3990	+	+	+	+	+	+		+	+	+	+	+
4000	+	+	+	+			+	+	+	+	+	+
4010	+	+	+	+	+	+	+					+
4020	+	+	+	+	+	+	+	+	+	+	+	+
4030	+	+	+	+	+	+	+	+	+	+	+	+
4040	+	+	+	+	+	+	+					+
4050	+	+	+	+			+				+	
4060	+	+	+	+	+	+	+	+	+	+	+	+
4070	+	+	+	+	+	+						
4080	+	+					+					
4090	+	+	+	+	+	+	+	+	+	+	+	+
4100	+	+	+	+			+	+	+	+	+	+

Erum	Eca	Ech	Ama	Aph	WBm	WMel	Nsen	Rbel	Rcon	Rfel	Rpro	Pub
4110	+	+	+	+	+	+	+	+	+	+	+	+
4120			+									
4130	+	+	+	+	+	+	+	+	+	+	+	+
4140	+	+	+	+	+	+	+	+	+	+	+	+
4150	+	+	+	+	+	+	+	+	+	+	+	
4160	+	+	+	+	+	+	+	+	+	+	+	+
4170	+	+		+	+	+	+	+	+	+	+	+
4180	+	+	+	+	+	+		+	+	+		
4190	+	+	+	+	+	+		+	+	+	+	
4200	+	+	+	+	+	+	+	+	+	+	+	+
4210	+	+										
4211	+	+	+	+	+	+						+
4220	+	+	+	+	+	+	+	+	+	+	+	+
4230	+	+	+	+	+	+						
4240	+	+	+	+	+	+	+	+	+	+	+	+
4250	+	+	+	+	+	+	+					+
4260	+	+	+	+	+	+	+	+	+	+	+	+
4261	+	+	+	+	+	+	+	+	+	+		
4270	+	+	+	+	+	+	+	+	+	+	+	+
4280	+	+	+	+	+	+	+	+	+	+	+	+
4310	+	+	+	+	+	+	+	+	+	+	+	+
4330		+	+	+	+	+		+	+	+		+
4340	+											
4350	+	+	+						+	+	+	
4360	+	+	+	+	+	+	+					
4370	+	+	+	+	+	+	+	+	+	+	+	+
4390	+	+										
4400		+										
4410	+	+	+	+	+	+						
4420	+	+	+	+	+	+	+	+	+	+	+	+
4430	+	+	+	+	+	+	+	+	+	+	+	+
4460	+	+	+	+	+	+	+	+	+	+	+	
4470	+	+				+						
4480	+	+	+	+	+	+						+
4490	+	+	+	+	+	+		+	+	+	+	
4500	+	+	+	+	+	+	+	+	+	+	+	+
4510	+	+		+	+	+						
4520	+	+	+	+			+					
4530		+										
4540	+	+	+	+	+	+	+	+	+	+	+	+
4550	+	+	+	+	+	+	+	+	+	+	+	+
4560	+	+			+	+		+	+	+	+	
4570	+	+	+	+	+	+	+					+
4580	+	+	+	+								
4590	+	+	+	+	+	+	+	+	+	+	+	+
4600	+	+	+	+	+	+	+	+	+	+	+	+
4660	+	+	+	+		+	+					
4670	+	+	+	+	+	+						
4680	+	+	+	+	+	+						
4690	+	+	+	+	+	+	+	+	+	+	+	+
4700	+	+	+	+	+	+	+	+	+	+	+	+
4710	+	+	+	+		+	+					
4720	+	+	+	+	+	+	+	+	+	+	+	+
4730	+	+	+	+	+	+	+					+
4750	+	+	+	+	+	+	+					+
4760	+	+	+	+	+	+		+	+	+	+	+
4770	+	+	+	+	+	+	+	+	+	+	+	+
4780	+	+	+	+	+	+	+	+	+	+	+	+
4790	+	+	+	+	+	+	+	+	+	+	+	+
4800	+	+		+	+	+		+			+	+
4810	+	+	+	+	+	+	+	+	+	+	+	+
4820	+	+	+	+	+	+	+	+	+	+	+	+
4830	+	+	+	+	+	+	+	+	+	+	+	+
4840	+	+	+	+	+	+	+					+
4850	+	+	+	+	+	+	+	+	+	+	+	+
4860	+	+	+	+	+	+	+	+	+	+	+	+
4870	+	+	+	+	+	+	+	+	+	+	+	
4880	+	+	+	+	+	+	+	+	+	+		+
4890	+	+	+	+		+	+	+	+	+	+	+
4900	+	+	+	+	+	+	+	+	+		+	+
4910	+	+	+	+	+	+	+	+	+	+	+	+
4920	+	+	+	+	+	+						
4930	+	+										
4940	+	+	+	+	+	+	+	+	+	+	+	+
4950	+	+	+	+		+	+	+	+	+		
4970	+	+	+	+	+	+	+	+	+	+	+	

Erum	Eca	Ech	Ama	Aph	WBm	WMel	Nsen	Rbel	Rcon	Rfel	Rpro	Pub
4980	+	+	+	+			+					+
4990	+		+	+	+	+	+	+	+	+	+	+
5000	+	+	+									
5010	+	+										
5020	+	+	+	+	+	+	+	+	+	+	+	+
5030	+	+	+	+	+	+	+	+	+	+	+	+
5040	+	+	+	+	+	+	+	+	+	+	+	+
5050	+	+										
5060	+	+	+	+		+			+			
5070	+	+			+	+						
5080	+	+	+	+	+	+	+	+	+	+	+	+
5090	+	+	+	+	+	+	+	+	+	+	+	+
5100	+	+	+	+	+	+	+	+	+	+	+	
5110	+	+	+	+	+	+	+	+	+	+	+	+
5120	+	+	+	+		+						
5130	+	+	+	+	+	+	+	+	+	+	+	+
5140	+	+										
5150	+	+	+	+	+	+	+					
5160	+	+	+	+	+	+	+	+	+	+	+	
5170	+	+	+	+	+	+	+					+
5180	+	+	+	+	+	+	+					+
5190	+	+	+	+	+	+	+	+	+	+	+	+
5200	+	+	+	+		+		+	+	+		
5210	+	+	+	+	+	+	+					
5220	+	+	+	+								
5230	+	+	+	+								
5240	+	+	+	+	+	+	+					
5250	+	+	+	+	+	+	+	+	+	+	+	
5260	+	+	+	+	+	+	+	+	+	+		
5270	+	+	+	+	+	+	+	+	+	+	+	
5280	+	+	+	+	+	+	+					+
5290	+	+	+	+	+	+	+	+	+	+	+	+
5300	+	+										
5310		+										
5320	+	+	+	+	+	+	+	+	+	+	+	+
5330	+	+	+	+		+	+	+	+	+	+	+
5340	+	+										+
5350	+	+	+	+	+	+	+	+	+	+	+	+
5360	+	+	+	+	+	+	+	+	+	+	+	+
5370	+	+	+	+								
5380	+	+	+	+	+	+						+
5390	+	+	+	+	+	+	+	+	+	+	+	
5400	+	+	+	+	+	+	+	+	+	+	+	
5410	+	+	+	+	+	+	+					
5420	+	+	+	+	+	+	+	+	+	+	+	
5430	+	+	+	+	+	+	+	+	+	+	+	+
5440	+	+	+	+	+	+		+				
5470	+											
5490	+	+	+	+	+	+	+		+	+		
5500	+	+	+	+	+	+	+	+	+	+	+	+
5510	+	+	+	+	+	+	+	+	+	+	+	+
5520	+	+	+	+	+	+	+	+	+	+	+	
5530	+		+		+			+				
5540	+		+		+				+	+	+	
5550	+	+	+	+	+	+	+	+	+	+		
5560	+	+	+	+	+							
5600	+	+	+	+	+	+	+					+
5610	+	+	+	+	+	+				+		
5620	+	+	+	+	+	+		+	+	+	+	+
5630	+	+	+	+	+	+	+					+
5640	+	+	+	+	+	+		+	+	+	+	
5650	+	+	+	+	+	+	+	+	+	+	+	+
5660	+	+	+	+	+	+	+					+
5670	+	+	+	+	+	+						
5680	+	+	+	+								
5690	+	+	+	+	+	+	+	+	+	+	+	
5700	+	+	+	+	+							
5710	+	+	+	+	+	+	+	+	+	+	+	+
5720	+	+	+	+	+	+	+	+	+	+	+	
5730	+	+	+	+	+	+	+					+
5740	+	+	+	+	+	+	+	+	+	+	+	+
5750	+	+	+	+	+	+	+	+	+	+	+	+
5760	+	+	+	+	+	+	+	+	+	+	+	+
5770	+	+	+		+	+		+	+	+	+	+
5780	+	+	+	+			+	+	+	+	+	
5790	+	+	+	+	+	+	+	+	+	+	+	+

Erum	Eca	Ech	Ama	Aph	WBm	WMel	Nsen	Rbel	Rcon	Rfel	Rpro	Pub
5791	+	+	+	+	+	+	+	+	+	+	+	
5800	+	+	+	+								
5810	+	+	+	+	+	+	+					
5820	+	+	+	+		+			+			
5830	+	+	+	+	+	+	+	+	+	+	+	+
5840	+	+	+	+	+	+	+	+	+	+	+	+
5850	+	+	+	+	+	+	+	+	+	+	+	+
5860	+	+	+	+	+	+	+	+	+	+	+	+
5870	+	+	+	+	+	+	+	+	+	+	+	+
5880	+	+	+	+	+	+	+	+	+	+	+	+
5890	+	+	+	+	+	+	+	+	+	+	+	+
5900	+	+	+	+	+	+	+	+	+	+	+	+
5910	+	+	+	+	+	+	+	+	+	+	+	+
5920	+	+	+	+	+	+	+	+	+	+	+	+
5930	+	+	+	+	+	+	+	+	+	+	+	+
5940	+	+	+	+	+	+	+	+	+	+	+	+
5950	+	+	+	+	+	+	+	+	+	+	+	+
5960	+	+	+	+	+	+	+	+	+	+	+	+
5970	+	+	+	+	+	+		+	+	+	+	+
5980	+	+	+	+	+	+	+	+	+	+	+	+
5990	+	+	+	+	+	+	+	+	+	+	+	+
5991	+	+	+	+	+	+	+	+				
6000	+	+	+	+	+	+	+	+	+	+	+	+
6010	+	+	+	+	+	+	+	+	+	+	+	+
6020	+	+	+	+	+	+	+	+	+	+	+	+
6030	+	+	+	+	+	+	+	+	+	+	+	+
6040	+	+	+	+	+	+	+	+	+	+	+	+
6050	+	+	+	+	+	+		+		+	+	
6060	+	+	+	+	+	+	+	+	+	+	+	
6070	+	+	+	+	+	+	+	+	+	+	+	+
6080	+	+	+	+	+	+	+	+	+	+	+	+
6090	+	+	+	+	+	+	+	+	+	+	+	+
6100	+	+	+	+	+	+	+	+	+	+	+	+
6110	+	+	+	+				+	+	+	+	+
6120	+	+	+	+	+	+	+	+	+	+	+	+
6130	+	+	+	+			+	+	+	+	+	+
6140	+	+		+	+	+		+	+	+	+	
6150	+	+	+	+								
6160	+	+	+	+	+	+						
6170	+	+				+						
6180	+	+	+	+	+	+	+	+	+	+	+	+
6190	+	+	+	+				+	+	+	+	+
6200	+	+	+	+		+	+	+	+	+		
6210	+	+	+	+		+						
6220	+	+	+	+	+	+						
6230		+										
6250	+	+	+	+	+	+	+	+	+	+	+	
6260	+	+	+	+	+	+	+					+
6270	+	+	+	+	+	+	+	+	+	+	+	
6280	+	+					+	+	+	+		
6290			+	+		+						+
6300	+	+										
6310	+	+	+	+	+	+	+					+
6320	+	+		+								
6330	+	+	+	+	+	+	+	+	+	+	+	+
6340	+	+	+	+								+
6350	+	+	+	+	+	+	+					+
6360	+	+	+	+	+	+	+	+	+	+	+	+
6370	+	+	+	+	+	+	+					+
6380	+	+	+	+	+	+	+	+	+	+	+	+
6390	+	+	+	+	+	+						
6400	+	+	+	+	+	+	+	+	+	+	+	
6410	+	+	+	+	+	+	+					
6420	+	+	+	+	+	+	+	+	+	+	+	+
6430	+	+	+	+	+	+	+	+	+	+	+	+
6440	+	+	+	+	+	+	+	+	+	+		
6450	+	+	+	+	+	+	+					+
6460	+	+	+	+	+	+	+	+	+	+	+	+
6470	+	+	+	+	+	+	+					+
6480	+	+	+	+	+	+		+		+		
6490	+	+	+	+	+	+						
6500	+	+	+	+			+			+		
6510	+	+	+	+	+	+	+		+			
6520	+	+	+	+			+	+	+			
6530	+	+										
6540	+	+	+	+	+	+				+		

Erum	Eca	Ech	Ama	Aph	WBm	WMel	Nsen	Rbel	Rcon	Rfel	Rpro	Pub
6550	+	+	+	+	+	+	+	+	+	+	+	+
6560	+	+	+	+		+						
6570	+	+	+	+	+	+						
6580	+	+	+	+	+	+	+					+
6590	+	+	+	+	+	+	+	+	+	+	+	
6600	+	+	+	+	+	+	+	+			+	
6610	+	+	+	+				+	+	+	+	+
6620	+	+	+	+	+	+	+	+				
6640	+	+	+	+	+	+						+
6650	+	+	+	+	+	+	+	+	+	+	+	
6660	+	+	+	+	+	+	+	+	+	+	+	+
6670	+	+	+	+	+	+	+					
6680	+	+	+	+	+							
6690	+	+	+	+	+	+	+	+	+	+	+	+
6700	+	+	+	+	+	+	+	+	+	+	+	
6710	+	+	+	+		+	+	+	+	+	+	+
6720	+	+	+	+	+	+	+	+	+	+	+	+
6730	+	+	+	+	+	+	+	+	+	+	+	+
6740	+	+	+	+	+	+	+	+	+	+	+	+
6750	+	+	+	+	+	+	+	+	+	+	+	+
6760	+	+	+	+	+	+	+	+	+	+	+	
6770	+	+	+	+	+	+	+	+	+	+	+	
6780	+	+	+	+	+	+	+					
6790	+	+	+	+	+	+	+	+	+	+	+	+
6800	+	+	+	+	+	+	+	+	+	+	+	+
6810	+	+	+	+	+	+	+	+	+	+	+	+
6820	+	+	+	+	+	+	+	+				
6830	+	+	+	+								
6840	+	+	+	+	+	+	+	+	+	+	+	+
6850	+	+	+	+	+	+	+	+	+	+	+	+
6860	+	+	+	+	+	+	+	+	+	+	+	+
6870	+	+	+	+	+	+		+	+	+	+	+
6880	+	+	+		+	+						
6890	+	+	+	+			+					
6900	+	+	+			+		+	+	+	+	
6910	+	+	+	+								+
6920	+	+	+	+		+	+	+	+	+	+	+
6930	+	+	+	+	+	+	+		+	+	+	+
6940	+	+	+	+	+	+	+	+	+	+	+	+
6950	+	+	+	+	+	+			+	+	+	
6960	+	+	+	+				+	+	+	+	+
6970	+	+	+	+	+	+		+	+	+	+	+
6980	+	+	+		+	+						+
6990	+	+	+	+	+	+	+	+	+	+	+	+
7000	+	+	+	+	+	+	+	+	+	+	+	+
7010	+	+	+	+	+	+	+	+	+	+	+	
7030	+	+	+	+	+	+	+					
7040	+	+	+	+	+	+	+	+	+	+	+	+
7050	+	+	+	+	+	+	+	+	+	+	+	
7170	+	+			+	+		+	+	+	+	
7180	+	+										
7220	+	+	+	+	+	+	+					
7230	+	+	+	+	+	+	+	+	+	+	+	+
7240	+	+	+	+	+	+	+	+	+	+	+	
7250	+	+	+	+								
7260	+	+	+	+	+	+	+	+	+	+	+	+
7270	+											
7290	+	+	+	+				+	+	+	+	
7390	+	+	+	+	+	+	+					+
7400	+	+	+	+	+	+	+					
7410	+	+	+	+	+	+	+	+	+	+	+	
7420	+	+	+	+	+	+	+	+	+	+	+	+
7430	+	+	+	+	+	+	+	+	+	+	+	
7440	+	+	+	+	+	+	+					+
7450	+	+	+	+	+	+	+					
7460	+	+	+	+	+	+	+	+	+	+	+	+
7470	+	+	+	+	+	+	+	+	+	+	+	+
7480	+	+	+	+	+	+	+					+
7490	+	+	+	+	+	+	+	+	+	+	+	+
7500	+	+	+	+	+	+	+	+				+
7510	+	+	+	+			+	+	+	+	+	
7520	+	+	+	+	+	+	+	+	+	+	+	
7530	+	+	+	+	+	+	+	+	+	+	+	
7540	+	+	+	+	+	+	+	+	+	+	+	+
7550	+	+			+	+	+	+	+	+	+	
7560	+	+	+	+				+	+	+	+	

Erum	Eca	Ech	Ama	Aph	WBm	WMel	Nsen	Rbel	Rcon	Rfel	Rpro	Pub
7570	+	+	+	+	+	+	+					+
7580	+	+	+	+	+	+	+	+	+	+	+	
7590	+	+	+	+	+	+	+	+		+		+
7600	+	+										
7610	+	+	+	+	+	+	+	+	+	+	+	
7620	+	+										
7630	+	+	+	+			+					+
7640	+	+	+	+								
7650	+	+										
7660	+	+	+	+	+	+	+	+	+	+	+	+
7661	+	+	+	+		+	+					
7670	+	+	+	+								
7680	+	+	+	+	+	+	+	+	+	+	+	+
7690	+	+	+	+	+	+	+	+	+	+	+	+
7700	+	+	+	+	+	+	+	+	+	+	+	+
7710	+	+	+	+	+	+	+	+	+	+	+	+
7720	+	+	+		+	+		+	+	+	+	+
7730	+	+	+	+	+	+	+	+	+	+	+	+
7740	+	+	+	+	+	+	+	+	+	+	+	+
7750	+	+	+	+	+	+	+	+	+	+	+	+
7760	+	+	+	+	+	+		+	+	+	+	+
7770	+	+	+	+	+	+	+					+
7780	+	+	+	+	+	+	+	+	+	+	+	+
7790	+	+						+	+	+	+	+
7800	+	+	+	+	+	+						
7810	+	+	+	+	+	+	+	+	+	+	+	+
7820	+	+	+	+	+	+	+	+	+	+	+	+
7830	+	+										+
7840	+	+	+	+	+	+	+	+	+	+	+	
7850	+	+	+	+			+	+				
7860	+	+	+	+	+	+	+	+	+	+	+	
7870	+	+	+	+	+	+	+	+	+	+	+	+
7880	+	+	+	+	+	+	+	+	+	+	+	+
7890	+	+	+	+	+	+	+	+	+	+	+	+
7900	+	+	+	+	+	+	+	+	+	+		+
7910	+	+	+	+	+	+						+
7920	+	+	+	+	+	+	+	+	+	+	+	+
7930	+	+	+	+	+	+						
7940	+	+	+	+	+	+	+					+
7950	+	+										
7960	+	+										
7970	+	+	+									
7980	+	+	+	+	+	+						
7990			+	+								
8000			+									
8010	+	+				+						
8020	+	+		+								
8030	+	+	+	+	+	+	+	+	+	+	+	+
8040	+	+	+	+	+	+	+	+	+	+	+	+
8050	+	+	+	+	+	+	+	+	+	+	+	+
8060	+	+	+									
8070	+	+	+	+	+	+	+	+	+	+	+	+
8080	+	+	+	+	+	+	+	+	+	+	+	+
8090	+	+	+	+	+	+	+					
8100	+	+	+	+	+	+						
8120	+	+	+	+	+	+	+	+	+	+	+	+
8130	+	+	+	+	+	+	+					+
8140	+	+	+	+	+	+	+	+	+	+	+	+
8150	+	+	+	+	+	+	+					
8160	+	+	+	+	+	+	+	+	+	+	+	+
8200	+	+	+	+	+	+	+	+	+	+	+	+
8210	+	+	+	+	+	+	+	+	+	+	+	+
8220	+	+	+	+	+	+	+	+	+	+	+	+
8230	+	+	+	+								
8240	+	+	+	+	+	+		+	+	+	+	
8250	+	+	+	+	+	+	+	+	+	+	+	+
8260	+	+	+	+	+	+			+	+		
8270	+	+	+	+								
8280	+	+	+	+	+	+	+	+	+	+	+	+
8290	+	+	+	+	+	+	+					+
8300	+	+	+	+	+	+	+	+	+	+	+	+
8310	+	+	+	+		+	+					
8320	+	+	+	+	+	+		+		+		
8330	+	+	+	+	+	+						+
8350	+	+	+	+		+	+					
8360	+	+	+	+	+	+	+	+	+	+	+	+

Erum	Eca	Ech	Ama	Aph	WBm	WMel	Nsen	Rbel	Rcon	Rfel	Rpro	Pub
8370	+	+	+	+	+	+	+	+	+	+	+	+
8380	+	+		+	+							
8390	+	+										
8400	+	+	+	+	+	+		+	+	+	+	
8410	+	+	+	+								+
8420	+	+	+	+	+	+		+		+	+	+
8430	+	+	+	+	+	+	+	+	+	+	+	+
8440	+	+	+	+	+	+	+	+	+	+	+	+
8450	+	+										
8460	+	+		+	+	+	+	+	+	+		
8470	+	+	+	+	+	+	+	+	+	+	+	+
8480	+	+	+	+			+					+
8490	+	+	+	+	+	+	+					+
8500	+	+	+	+	+	+	+	+	+	+	+	+
8510	+	+	+	+								
8520	+	+	+	+	+	+	+	+	+	+	+	+
8530	+	+	+	+	+	+	+	+	+	+	+	+
8550	+	+	+	+	+	+	+	+	+	+	+	+
8560	+	+	+	+	+	+	+	+	+	+	+	+
8570	+	+	+	+	+	+	+	+	+	+	+	+
8580	+	+	+	+			+	+	+	+	+	+
8590	+	+										
8600	+	+										
8620	+	+										
8630	+	+										
8640	+	+										
8650	+	+										
8660	+	+										
8710	+	+										
8730	+	+		+	+					+		
8740	+	+	+		+							
8750	+	+										
8770	+	+										
8780	+	+	+	+	+	+	+		+	+	+	+
8790	+	+										
8800	+	+	+	+	+	+	+	+	+	+	+	+
8810	+	+		+		+						+
8820	+	+	+	+	+	+	+		+	+	+	+
8830	+	+	+			+		+	+	+	+	+
8840	+	+	+			+	+	+	+	+	+	+
8850	+	+	+	+	+	+		+	+	+	+	
8860	+	+	+	+	+	+	+	+	+	+	+	+
8870	+	+	+	+	+	+	+	+	+	+	+	+
8880	+	+	+	+								+
8890	+	+	+	+	+	+	+	+	+	+	+	+
8900	+	+	+	+	+	+	+	+	+	+	+	+
8910	+	+	+	+			+	+	+	+	+	+
8920	+	+	+	+	+	+	+	+	+	+	+	+
8930	+	+										

CHAPTER 4

Repetitive DNA in the complete genome sequence of *Ehrlichia ruminantium* (Welgevonden)

4.1. INTRODUCTION

DNA repeats can be defined as sequences sharing extensive similarity with other sequences in the same genome. Repetitive DNA can be divided into two main categories, dispersed repeat motifs and tandemly repeated sequences. Dispersed repeats are either in the same orientation as direct repeats, or they can occur in reverse orientation on opposite strands of the chromosome. Some repeat units are located close together, but they can be situated kilobases apart. Tandem repeats consist of either simple homopolymeric tracts of a single nucleotide or of multimeric repeats. These multimeric repeats are built from identical units (homogeneous repeats), mixed units (heterogeneous repeats), or degenerate repeat sequence motifs (Van Belkum *et al.*, 1998). During the annotation of the *E. ruminantium* genome we labelled tandem repetitive regions as "repeat regions" (identification codes of the form rptnnn) and dispersed repeats as "repeat units" (identification codes of the form rpt_unit_nnn). For ease of reference throughout the rest of this chapter note that Tables 4.2 and 4.4 list the identification codes of the repeat regions and repeat units, respectively.

Illegitimate recombination can occur between tandem repeats, or repeats located close together, through slipped-strand mispairing at replication pauses or single strand annealing following exonucleolytic degradation at a DNA double-strand break (Levinson & Gutman, 1987; Rocha, 2003). The effects of such recombination events may not result in major chromosomal rearrangements, but if an event occurs within a gene it can change the coding frame of the gene (phase variation), and in surface antigens it could affect antigenicity. If the illegitimate recombination event occurs in a non-coding region it may have an effect on the expression of nearby genes by disrupting promoter sequences.

DNA repeats can be used by the RecA protein to repair damaged chromosomes by using a duplicate copy of the damaged sequence as a template for repair (Hughes, 2000a). In the repair process homologous recombination can take place, which can result in rearrangements of genes or parts of genes, tandem duplications, translocations and inversions.

In the field of genetics, use is frequently made of shorter tandem repeats as molecular markers (Nakamura *et al.*, 1987), and it has even been proposed that short tandem repeats might identify putative virulence genes (Hood *et al.*, 1996). There are many examples where immunoreactive bacterial proteins are found to contain repeats and *Ehrlichia* species provide several particular instances. For example, a subset of tandem repeat-containing proteins that elicit strong host immune responses and are associated with host-pathogen interactions has been identified in both *E. chaffeensis* and *E. canis* (Luo *et al.*, 2008).

The *E. ruminantium* genome sequence contains unusually large amounts of repetitive DNA (Table 2.1, Figure 2.2). In this chapter these repeats will be discussed in detail and compared with repetitive sequences identified in the genome sequences of other members of the order Rickettsiales.

4.2. MATERIALS AND METHODS

See Appendix B for materials and media components.

4.2.1. Analysis of genomic repeat sequences

During the analysis of the *E. ruminantium* genome sequence (sub-section 2.2.2) mreps (Kolpakov *et al.*, 2003) and Tandem Repeats Finder (Benson, 1999) were used to locate tandem repeats, while GAP4 (Bonfield *et al.*, 1995) and Dotter (Sonnhammer & Durbin, 1995) were used to identify dispersed repeats. Ankyrin repeat domains were identified with Pfam (Bateman *et al.*, 2004).

Tandem repeats were also identified in the genome sequences of the other members in the order Rickettsiales using Tandem Repeats Finder. To simplify the comparisons we used the default parameters of Tandem Repeats Finder for all searches. The search results were converted to a format that can be visualised in Artemis (Rutherford *et al.*, 2000) and the ACT program (Carver *et al.*, 2005).

The organisms included in the analysis are listed in Table 3.1 and discussed in section 3.1. The complete genome sequences were retrieved and aligned as described in sub-section 3.2.2.

4.2.2. Amplification and cloning of variable repeat regions

The following primers were used to amplify the regions containing variable numbers of tandem repeat units: 758_RC1_F and 758_RC1_F (rpt121, Table 4.2); 758_RC2_F and 758_RC2_F (rpt148, Table 4.2); WTHIN440_5F and WTHIN440_5R (rpt18, Table 4.2); and WGAP71walk_1F and WGAP71R (rpt110, Table 4.2). The sequences of these primers can be found in Appendix C1. Template genomic DNA was prepared as described in sub-section 2.2.1.1 and PCR amplifications were conducted using the Platinum® *pfx* DNA polymerase kit (Invitrogen). Each 50 µl reaction contained 25 ng genomic DNA, *pfx* PCR buffer, 0.3 mM

dNTPs, 1 mM $MgSO_4$, 0.2 µM of each primer and 1 U *pfx* DNA polymerase. The reaction conditions consisted of one cycle of 5 min at 94°C, 30 cycles of 20 s at 94°C, 30 s at 50°C and 2 min at 68°C, followed by a final incubation of 10 min at 68°C. Amplified products were visualised by electrophoresis on a 1% agarose gel, stained with ethidium bromide. The amplicons were purified with the High Pure PCR Product Purification kit (Roche) and cloned into the pGEM-T Easy vector (Promega) using the protocols provided by the manufacturers. Plasmid DNA was isolated using the High Pure Plasmid Isolation kit (Roche) according to the manufacturer's instructions and digested with *Eco*RI (Roche). The inserts were visualised on 1% agarose gels. At least 20 clones of each region were selected and sequenced with the SP6 and T7 primers (Appendix C4). We sequenced several clones fourfold to show that any observed variation was not an artefact of the sequencing process. Sequencing reaction conditions were as described in sub-section 2.2.1.4.

4.2.3. Amplification of the regions around the *rho* and *tuf* genes

In this part of the investigation we used genomic DNA from all the *E. ruminantium* isolates that were available in our laboratory at the time of the investigation (Figure 4.2). Primers were designed to amplify the *tuf* and *rho* regions; the sequences of the primers can be found in Appendix C2. The combinations of *rho* primers used in each reaction are illustrated in Figure 4.2 and the same procedure was followed to investigate the *tuf* regions. The PCR reactions contained 25 ng genomic DNA, 0.25 µM of each primer, 0.2 mM dNTPs and 1 U TaKaRa Ex Taq™ (TaKaRa Bio Inc.). The reaction conditions were: one cycle of 5 min at 94°C, 30 cycles of 10 s at 94°C, 30 s at 52°C and 4 min at 72°C, and a final extension of 7 min at 72°C. Amplified products were analysed by electrophoresis on 1% agarose gels containing ethidium bromide.

4.3. RESULTS AND DISCUSSION

4.3.1. Repeat sequences in the *E. ruminantium* genome sequence

One of the most striking features of the *E. ruminantium* genome is the large number of tandem repeats and dispersed repeated sequences, including both direct and inverted repeats. These constitute 8.5% of the chromosome and contribute to the high proportion of non-coding sequence, which results in a larger size for the *E. ruminantium* genome than for most other Rickettsiales (Table 4.1). The *E. ruminantium* genome contains more tandem repeats (158) than any of the other members in the order Rickettsiales, followed by *E. chaffeensis* with 125 repeats. The biggest genome in the order, that of *R. bellii* (1.522 Mb), also contains a fairly large number of tandem repeats (79). By contrast, very few repeated sequences were identified in the two smallest genomes, the 0.859 Mb genome of *Neorickettsia sennetsu* (13) and the 1.08 Mb *Wolbachia pipientis w*Bm genome (11). The *W. pipientis w*Mel genome, on the other hand, is also relatively small (1.27 Mb) but contains large numbers of DNA repeats (81). The irregular GC-skew pattern in *W. pipientis w*Mel has been attributed to intragenomic rearrangements associated with the repeat elements (Wu *et al.*, 2004). However, the typical GC-skew pattern seen in many other bacteria, with transitions in GC-skew values at the origin and termination of replication, is maintained in the repeat-rich *E. ruminantium* genome. Interestingly the free-living bacterium *Pelagibacter ubique* also contains a rather high number of tandem repeats (71), which is particularly surprising as *P. ubique* has the smallest genome (1.3 Mb) of any known independently replicating cell. It is surmised that evolution has reduced this genome to the minimum size required for efficient growth in a nutritionally poor environment (Williams *et al.*, 2007), which would therefore suggest that the DNA repeats play a vital survival role. In this free-living organism that role cannot be related to the generation of variation in immunoreactive surface proteins, which is normally assumed to be an important function of repeats in parasitic bacteria (see sub-section 4.3.3.1).

Table 4.1. Genome properties of the sequenced genomes in the order Rickettsiales.

Family	Species	Genome size (Mb)	% GC	Number of CDS	Average CDS length (bp)	% coding	Number of tandem repeats
Anaplasmataceae	*E. ruminantium*[1]	1.516	27.48	920	1032	62.0	158
	E. canis[2]	1.315	28.96	925	1025	72.1	75
	E. chaffeensis[3]	1.176	30.09	1104	847	79.5	125
	Anaplasma marginale[4]	1.198	49.76	949	1081	85.7	47
	A. phagocytophilum[3]	1.471	41.64	1263	794	68.1	76
	Neorickettsia sennetsu[3]	0.859	41.08	931	808	87.6	13
	*Wolbachia pipientis w*Mel[5]	1.268	35.23	1195	850	80.2	81
	*W. pipientis w*Bm[6]	1.080	34.18	805	899	67.0	11
Rickettsiaceae	*Rickettsia bellii*[7]	1.522	31.65	1428	907	85.1	79
	R. conorii[8]	1.269	32.44	1373	746	80.7	29
	R. felis[9]	1.587	32.45	1399	889	83.7	65
	R. prowazekii[10]	1.112	28.99	834	1006	75.5	32
SAR11 cluster	*Pelagibacter ubique*[11]	1.309	29.68	1354	925	95.7	71

[1]Collins *et al*., 2005; [2]Mavromatis *et al*., 2006; [3]Hotopp *et al*., 2006; [4]Brayton *et al*., 2005; [5]Wu *et al*., 2004; [6]Foster *et al*., 2005; [7]Ogata *et al*., 2006; [8]Ogata *et al*., 2000; [9]Ogata *et al*., 2005; [10]Andersson *et al*., 1998; [11]Giovannoni *et al*., 2005

4.3.2. Simple sequence repeats (SSRs)

One hundred and twenty-six SSRs of 1-5 bp were identified using mreps. Polymorphic homopolymeric tracts (usually of G or C nucleotides) and short repeats (2-5 bp) have been implicated in phase variation of surface-associated proteins in other bacteria (Parkhill *et al.*, 2000). In the *E. ruminantium* genome there were only four polymeric tracts of G or C nucleotides and only one of these was located within a gene. Only one was found to be polymorphic, C(11-12), but it was located in a non-coding region 622 bp from the start of the nearest gene. Several polymeric tracts of T or A nucleotides were identified, but again only one of these was polymorphic and it was also located far from the nearest start codon. Various other SSRs of 2-5 bp were identified, many of which were AT rich and located in intergenic regions. Thirteen SSRs were located within the promoter regions upstream of the predicted start codons of genes while only three were located within ORFs close to the start codons. Whether these SSRs play a role in promoter regulation or phase variation in the *E. ruminantium* genome remains to be elucidated.

4.3.3. Longer tandem repeats (LTRs)

Numerous LTRs (six bp up to 471 bp) were identified in the *E. ruminantium* genome sequence (Table 4.2). Five LTRs overlapped the 5' end of a gene and 20 the 3' end of a gene, while two overlapped the 3' end of one gene and the 5' end of the following gene. The majority (53.8%) of LTRs were located in non-coding regions, whereas 31.6% of LTRs occurred within genes. LTRs which overlap at the beginnings or ends of genes account for eight (25.0%) of the pseudogenes identified in *E. ruminantium*. In these cases the beginning or the end of a gene has been duplicated, producing a putative pseudogene. This has occurred four times, each time producing two pseudogenes.

Table 4.2. Tandem repeats in the *E. ruminantium* genome. (Adapted from Collins *et al.,* 2005. [Supplementary information])

ID code	Location of region (Co-ordinates)	Length of repeated motif (bp)	No. of units in region	Feature overlapping repeat region or within which region is located
rpt1	4449..4900	203	2.2	
rpt2	11386..11416	12	2.6	
rpt3	12752..13197	158	2.8	3' end of Erum0080
rpt4	29304..30133	99	8.4	Erum0250
rpt5	29304..30133	297	2.8	Erum0250
rpt6	31558..31587	6	5	Erum0260
rpt7	34831..34884	6	9	Erum0280
rpt8	46434..46947	151	3.4	
rpt9	48545..49149	203	3	5' end of Erum0370, Erum0371, Erum0372
rpt10	54146..54823	240	2.8	
rpt11	57994..58529	255	2.1	
rpt12	60821..61714	283	3.2	3' end of Erum0430, rpt_unit_3A-C
rpt13	68653..69240	198	3	3' end of Erum0490, Erum0841, Erum0842
rpt14	106486..107266	300	2.6	Erum0660
rpt15	106721..107991	471	2.7	Erum0660
rpt16	107021..107437	171	2.4	Erum0660
rpt17	107492..107908	171	2.4	Erum0660
rpt18	124367..124609	7	34.7	
rpt19	126403..127088	170	4	3' end of Erum0740
rpt20	134617..135028	137	3	
rpt21	137634..138614	336	2.9	3' end of Erum0780, Erum0781, Erum0782
rpt22	149410..149768	148	2.4	
rpt23	156223..156693	119	4	
rpt24	160963..161810	313	2.7	
rpt25	160998..161810	156	5.2	
rpt26	166158..166649	152	3.3	
rpt27	179073..179850	154	5.1	3' end of Erum1020
rpt28	183561..184359	294	2.8	Erum1040
rpt29	192068..192097	12	2.5	Erum1110, rpt_unit_8B
rpt30	192336..193846	27	56	Erum1110
rpt31	198548..199104	137	4.1	
rpt32	214942..215343	190	2.1	
rpt33	218937..219507	237	2.4	Erum1230
rpt34	243765..244327	203	2.8	rpt_unit_10A
rpt35	247950..248408	198	2.3	Erum1430
rpt36	272236..272683	149	3	
rpt37	296093..296823	251	2.9	
rpt38	299415..300161	208	3.6	3' end of Erum1760
rpt39	314304..314707	202	2	5' end of Erum1830
rpt40	349552..349581	15	2	
rpt41	358072..358255	45	4.1	Erum2090
rpt42	358077..358250	15	11.6	Erum2090
rpt43	358101..358232	30	4.4	Erum2090
rpt44	367354..367911	195	2.9	
rpt45	373565..374242	252	2.7	Erum2170
rpt46	373608..374244	126	5.1	Erum2170
rpt47	391766..392582	165	5	
rpt48	411844..412019	90	2	Erum2400
rpt49	438192..438658	155	3	3' end of Erum2530
rpt50	443734..444162	179	2.4	
rpt51	444240..444277	20	1.9	

ID code	Location of region (Co-ordinates)	Length of repeated motif (bp)	No. of units in region	Feature overlapping repeat region or within which region is located
rpt52	447350..447791	221	2	3' end of Erum2610
rpt53	452065..452850	187	4.2	3' end of Erum2630
rpt54	452065..452850	375	2.1	3' end of Erum2630
rpt55	456431..457015	165	3.6	
rpt56	473389..473985	149	4	
rpt57	475910..476473	151	3.7	
rpt58	489760..489800	21	2	Erum2780
rpt59	493722..493751	15	2	Erum2800
rpt60	515332..516192	144	6	5' end of Erum2950
rpt61	530289..530828	180	3	
rpt62	548431..549229	182	4.4	3' end of Erum3180, Erum3171, Erum3172
rpt63	566548..566582	18	1.9	
rpt64	571014..571911	134	6.7	
rpt65	574204..574814	242	2.5	
rpt66	574287..574653	117	3.1	
rpt67	619459..619496	12	3.2	Erum3570
rpt68	622191..622525	45	7.4	Erum3590
rpt69	622515..622601	42	2.1	Erum3590
rpt70	624642..624841	12	16.7	Erum3600
rpt71	624720..624835	6	19.3	Erum3600
rpt72	652889..652951	27	2.2	Erum3730
rpt73	654790..655014	27	8.3	Erum3750
rpt74	655492..656048	144	3.9	Erum3750
rpt75	698790..699173	144	2.7	Erum3980
rpt76	699239..699336	36	2.7	Erum3980
rpt77	699775..700630	93	9.2	Erum3980
rpt78	730104..730145	21	2	Erum4220
rpt79	758913..758945	16	2.1	
rpt80	779363..779405	22	2	Erum4530
rpt81	796148..796177	15	2	
rpt82	810633..811157	247	2.1	3' end of Erum4730
rpt83	811944..812898	138	6.9	Erum4740
rpt84	825306..825332	9	3	Erum4850
rpt85	853307..853356	24	2.1	Erum5010
rpt86	855095..855134	20	2	Erum5030
rpt87	864452..864507	9	6.2	
rpt88	871251..871901	214	3	
rpt89	877038..877671	179	3.5	
rpt90	877721..877752	16	2	
rpt91	881799..883129	261	5.1	Erum5210
rpt92	883692..884550	222	3.9	Erum5210
rpt93	884370..884720	180	1.9	Erum5210
rpt94	888684..889192	216	2.4	Erum5220
rpt95	892429..892463	15	2.3	Erum5220
rpt96	904475..905286	280	2.9	
rpt97	914552..914581	14	2.1	Erum5320
rpt98	918222..918736	129	4	
rpt99	921143..922460	140	9.5	
rpt100	921143..922460	279	4.7	
rpt101	929535..930044	207	2.5	
rpt102	932150..932880	186	3.9	
rpt103	941045..941870	189	4.4	
rpt104	956699..957238	183	3	Erum5570
rpt105	979430..980213	161	4.9	
rpt106	984613..985285	212	3.2	

ID code	Location of region (Co-ordinates)	Length of repeated motif (bp)	No. of units in region	Feature overlapping repeat region or within which region is located
rpt107	988179..988687	169	3	
rpt108	993513..993930	211	2	
rpt109	1002890..1004344	142	10.2	3' end of Erum5820
rpt110	1034462..1035245	122	6.5	
rpt111	1044574..1045287	132	5.4	3' end of Erum6250
rpt112	1057624..1058328	164	4.3	
rpt113	1065491..1066184	295	2.4	
rpt114	1073866..1074570	148	4.8	
rpt115	1085431..1086061	202	3.1	
rpt116	1095325..1095947	142	4.4	
rpt117	1099600..1100749	185	6.2	3' end of Erum6510
rpt118	1101733..1102283	124	4.4	3' end of Erum6520
rpt119	1110652..1111517	144	6	
rpt120	1111594..1111895	150	2	
rpt121	1114817..1115357	7	77.3	
rpt122	1125337..1126104	208	3.7	
rpt123	1138275..1139139	173	5	
rpt124	1143120..1143331	77	2.7	
rpt125	1149259..1150319	291	3.6	
rpt126	1158770..1159284	156	3.3	
rpt127	1171110..1172163	219	4.8	
rpt128	1175473..1176225	238	3.2	3' end of Erum6940
rpt129	1195479..1196223	183	4.1	
rpt130	1200254..1200804	141	3.9	Erum7070
rpt131	1201263..1202289	198	5.2	Erum7070
rpt132	1214924..1215664	142	5.2	
rpt133	1221582..1222079	178	2.8	3' end of Erum7170
rpt134	1229491..1229837	181	1.9	3' end of Erum7220
rpt135	1234057..1235052	137	7.3	
rpt136	1248149..1249038	226	3.9	
rpt137	1278505..1279305	237	3.3	
rpt138	1281565..1281989	212	2	
rpt139	1286740..1286771	10	3.2	
rpt140	1290658..1291359	99	7.1	
rpt141	1297593..1298323	149	4.9	
rpt142	1299322..1299351	16	1.9	Erum7600
rpt143	1321356..1322029	135	5	
rpt144	1347300..1347979	154	4.4	
rpt145	1352229..1352689	127	3.6	
rpt146	1360356..1360920	191	3	rpt_unit_71A
rpt147	1369576..1369615	15	2.7	Erum7960
rpt148	1396844..1397299	7	65.1	
rpt149	1403693..1403727	17	2.1	
rpt150	1439941..1440514	192	3	
rpt151	1450924..1452182	243	5.2	3' end of Erum8450
rpt152	1469598..1470035	146	3	
rpt153	1474500..1474526	13	2.1	
rpt154	1495833..1495984	24	6.3	Erum8770
rpt155	1495856..1495950	9	11.6	Erum8770
rpt156	1495865..1495961	15	7.7	Erum8770
rpt157	745887..745905	6	3.2	
rpt158	1475098..1475119	6	3.7	Erum8590

4.3.3.1. Tandem repeats in coding regions

Of the 31 CDSs containing LTRs, 27 (87.1%) are either genes whose products are predicted to be membrane-associated or are genes unique to *E. ruminantium* (Table 4.3). Examination of orthologous CDSs in all the Rickettsiales revealed that the orthologs do not contain homologs of the repeats identified in *E. ruminantium*. In contrast, the tandem repeats in CDSs in each *E. ruminantium* genome have identical homologs in orthologous CDSs in the other two *E. ruminantium* genomes (Frutos *et al.*, 2007). This suggests that the repeats were generated after *E. ruminantium* had split from the common ancestor of all *Ehrlichia* species.

Twenty-two of the 31 CDSs containing LTRs are larger than the average length for predicted *E. ruminantium* genes. They include Erum0660, Erum3750 and Erum3980 which are particularly large genes, predicted to encode proteins of 3715, 1674 and 3002 amino acids respectively. It is interesting to note that four of the genes coding for type IV secretion system proteins contain tandem repeats. The repeat motifs in two of these genes, *virD4* and *virB10*, were relatively short (6 bp motifs repeated five and nine times respectively), while those in the two large putative type IV secretion system proteins Erum5210 and Erum5220 were between 15 and 261 bp in length.

Erum1110 contains a 27 bp sequence motif that is repeated 56 times. Interestingly the upstream gene, Erum1100, appears to be a paralog of Erum1110; the first 382 bp of Erum1100 has 90.8% identity to the 5' end of Erum1110, but terminates where the repeat starts in Erum1110, and therefore does not contain the tandem repeat (Figure 4.3A). These genes will be discussed in more detail in sub-section 4.3.4.2. A gene homologous to Erum1110 but containing 21.7 copies of the 27 bp motif was previously identified in *E. ruminantium* (Highway) by immune screening of an expression library (Barbet *et al.*, 2001). A synthetic peptide containing the repeat was recognised in an ELISA assay by immune sera from *E. ruminantium*-infected animals, indicating that this gene codes for a protein which is recognised by the immune system of the host.

Pathogenic bacteria have on average higher densities of tandem repeats than their free-living counterparts (Rocha, 2003), which may be related to generating sequence variation in genes involved in pathogenesis and evasion of the host immune response, and the likely recognition of Erum1110 by the host immune system is in accord with this suggestion. Many immunodominant proteins from pathogenic bacteria contain such tandem repeats, including the major surface protein 1 (*msp1α*) from *Anaplasma marginale* in which a neutralisation sensitive epitope is present within each repeat unit (Allred *et al.*, 1990). *Mycoplasma hyorhinis* possesses a complex system of variable surface lipoproteins (Vlps) that can alter susceptibility to inhibition by host antibodies. The only difference between the allelic forms of Vlp size variants expressed on susceptible and resistant organisms is the number of internal repeat units in the 3' region of the genes. There appears to have been selection for Vlps containing a greater number of tandem repeats; it was suggested that the larger size of such proteins might provide a protective shield for other surface proteins that are less free to change (Citti *et al.*, 1997). Therefore, although proteins containing such repeats may have an essential role or impart selective advantage, they may not necessarily be useful vaccine targets.

Table 4.3. CDSs containing LTRs. (Adapted from Collins *et al.,* 2005. [Supplementary information])

Systematic ID	Length of ORF (bp)	Length of repeated motif (bp)	Frequency of repeat	ID code of repeat	Putative product
Erum0250	1374	297	2.8	rpt5	Unknown
Erum0260	2406	6	5	rpt6	type IV secretion system protein VirD4
Erum0280	1347	6	9	rpt7	type IV secretion system protein VirB10
Erum0660	11148	300 471 171 171	2.6 2.7 2.4 2.4	rpt14 rpt15 rpt16 rpt17	Unknown
Erum1040	3498	294	2.8	rpt28	probable integral membrane protein
Erum1110	1986	12 27	2.5 56	rpt29 rpt30	Unknown
Erum1230	561	237	2.4	rpt33	Unknown
Erum1430	2856	198	2.3	rpt35	Unknown
Erum2090	2568	45	4.1	rpt41	putative cell division protein FstK
Erum2170	3222	252	2.7	rpt45	Unknown
Erum2400	1176	90	2	rpt48	probable membrane protein
Erum2780	1575	21	2	rpt58	probable membrane protein
Erum2800	1563	15	2	rpt59	probable membrane protein
Erum3570	1131	12	3.2	rpt67	probable integral membrane protein
Erum3590	1170	45 42	7.4 2.1	rpt68 rpt69	probable integral membrane protein
Erum3600	1758	12	16.7	rpt70	probable integral membrane protein
Erum3730	462	27	2.3	rpt72	Unknown
Erum3750	5025	27 144	8.3 3.9	rpt73 rpt74	unknown, contains 19 ankyrin repeat domains
Erum3980	9009	144 36 93	2.7 2.7 9.2	rpt75 rpt76 rpt77	unknown, contains 7 ankyrin repeat domains
Erum4220	1539	21	2	rpt78	lysyl-tRNA synthetase
Erum4530	600	22	2	rpt80	Unknown
Erum4740	1920	138	6.9	rpt83	probable exported protein
Erum4850	1023	9	3	rpt84	conserved hypothetical GTP-binding protein
Erum5010	1695	24	2.1	rpt85	probable exported protein
Erum5030	1227	20	2	rpt86	cytochrome b
Erum5210	7368	261 222 180	5.1 3.9 1.9	rpt91 rpt92 rpt93	putative type IV secretion system protein
Erum5220	4590	216 15	2.4 2.3	rpt94 rpt95	putative type IV secretion system protein
Erum5320	1983	14	2.1	rpt97	probable acetyl-/propionyl-coenzyme A carboxylase alpha chain
Erum5570	1659	183	3	rpt104	Unknown
Erum7070	4122	141 198	3.9 5.2	rpt130 rpt131	probable membrane protein
Erum7960	2208	15	2.7	rpt147	unknown, contains a GTP-binding domain
Erum8590	930	6	3.7	rpt158	putative outer membrane protein MAP1-14
Erum8770	534	24	6.3	rpt154	Unknown

4.3.3.2. Repeat regions with variable number of repeat units

We were not able to obtain sufficient amounts of pure *E. ruminantium* DNA for genomic library construction from a single tissue culture flask, hence the DNA used to generate the libraries was obtained from several passages, representing many generations of the organism. It might have been expected, therefore, that the generation of tandem repeats by slipped-strand mispairing would have led to instances of variations in the numbers of repeats between different clones originating from different generations, and we did indeed identify four sites where there were variable numbers of repeats. We confirmed that the variation was not caused by PCR or sequencing artefacts by amplifying the repeat regions with a high-fidelity proof-reading polymerase (Figure 4.1) and sequencing several clones, including clones from the WL1 and WL3 libraries, four times. Interestingly, three of the instances involve tandem repeats of different 7 bp motifs, with markedly variable numbers (rpt121, 4-80; rpt148, 7-88, and rpt18, 16-38) of the repeated sequence motif. The fourth instance is a 122 bp repeat (rpt110) which occurs with continuously variable frequency from 1.5 to 7.5 times. When we amplified these repeat regions each of the 7 bp repeat amplicons appeared to be a single band of distinct size (Figure 4.1, Panel A: 7 bp repeat 1-3). However, the clones of the amplicons contained inserts of varying lengths (Figure 4.1, Panel B), suggesting that the variation in the number of motifs could be the result of cloning the amplicons into *E. coli*. Unfortunately it was impossible to sequence through the repeats directly from the PCR product. Hence, it is still unclear whether the 7 bp repeat regions in fact contain a variable number of repeat units or whether it is the *E. coli* host cell that cannot maintain the original numbers of repeats. In contrast the different sizes of amplified repeat units for the 122 bp repeat were clearly visible in its PCR product (Figure 4.1, Panel A).

Of the three 7 bp repeat regions one (rpt18) cannot be translated into ORFs, another (rpt121) can be translated on all three forward frames, and the third (rpt148) has ORFs in all six frames. However, none of the translated ORFs are predicted to be protein-coding. All three 7 bp tandem repeat regions have a higher G+C content than the rest of the genome and exhibit strand asymmetry (one strand contains predominantly either Gs or Cs). Other G+C rich hypervariable

sequences have been shown to form secondary structures, which can cause DNA polymerase to pause and may result in the rapid generation of tandem repeats (Weitzmann *et al.*, 1997). The formation of secondary structures thus may explain the variability in the number of these 7 bp repeat units.

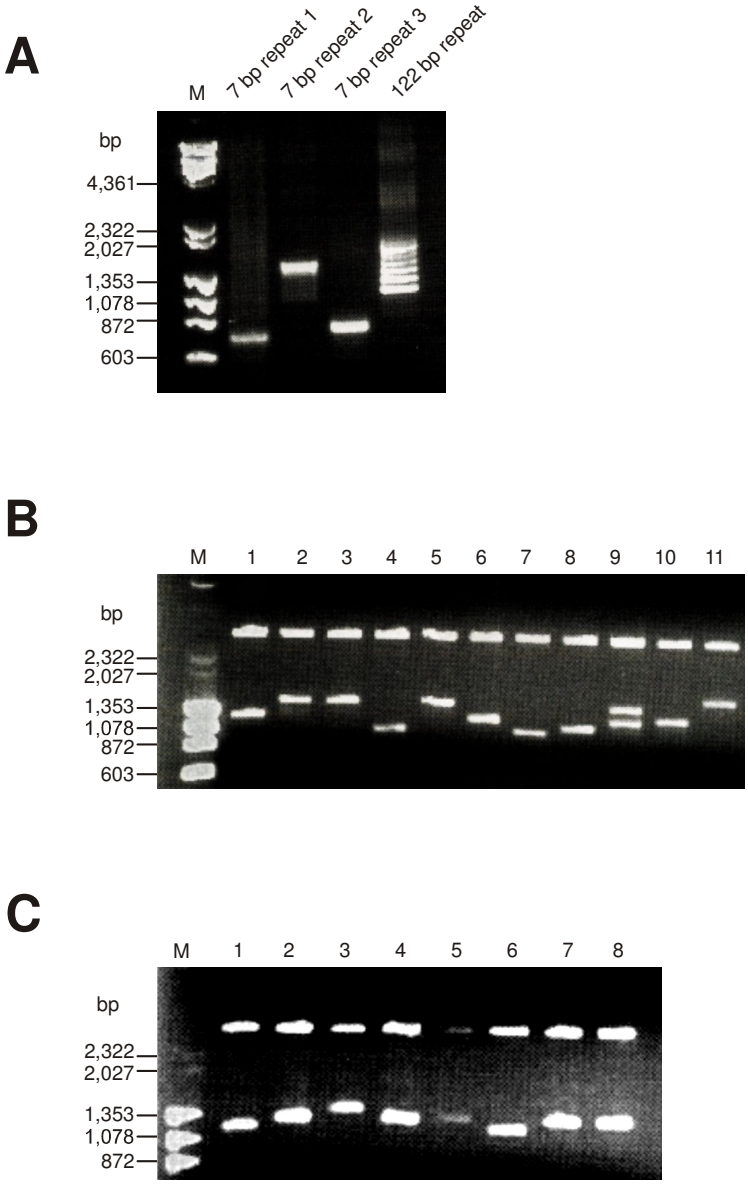


Figure 4.1. Amplification and cloning of variable repeat regions from *E. ruminantium* Welgevonden genomic DNA. **A**. Repeat regions rpt121 (7 bp repeat 1), rpt148 (7 bp repeat 2), rpt18 (7 bp repeat 3), and rpt110 (122 bp repeat) amplified by PCR. **B.** Some of 7 bp repeat 2 (rpt148) clones showing a large variation in insert size. **C**. Clones of the 122 bp repeat (rpt110). Lambda *Hin*dIII combined with ΦX174 *Hae*III markers are in lanes labelled M.

4.3.4. Interspersed repetitive DNA

There were numerous duplicated sequences in the genome, including both direct and inverted repeats (Table 4.4). We identified 75 such repeat units, the majority of which were present twice in the genome; there were three copies of four of the repeat units and four copies of two. The repeat units ranged in size from 64 bp to almost 3 kb, with the majority between 100 and 400 bp; repeat units were from 75% - 100% identical. Approximately equal numbers of direct and inverted repeat units were identified. Translocation and inversion events have resulted in the duplication and truncation of a number of genes; in fact, 21 (65.6%) of the putative pseudogenes that were identified appear to have been produced in this way. We identified five large duplications (> 1 kb) in the genome, four of these were associated with genes, and one was located in an intergenic region.

4.3.4.1. Homologous recombination between repetitive sequences

Both chromosomal inversions and translocations are common between closely related species and inversions are frequently symmetrical around the origin of replication. These inversions occur between repeated sequences and result in a reversal of the genomic sequence between the repeats (Hughes, 2000b). As described in section 4.1 these chromosomal rearrangements are often the result of RecA mediated repair of damaged DNA.

Another consequence of RecA mediated homologous recombination is gene conversion (Wiuf & Hein, 2000; Chen *et al.*, 2007). This mechanism involves the unidirectional transfer of genetic material from one region to the corresponding place in another paralogous region, which results in homogenisation of the sequences of repeated genes (Petes & Hill, 1988; Lawson *et al.*, 2009; Osada & Innan, 2009). In *E. ruminantium*, gene conversion appears to have limited the divergence between the *rho1* and *rho2* and the *tufA* and *tufB* genes which respectively have 94.0% and 100% identity in overlapping regions. In *Salmonella enterica* serovar Typhimurium such co-evolution of the *tufA* and *tufB* genes has been linked to chromosomal rearrangements (Hughes,

2000b). Both the *rho* and the *tuf* genes are in inverse order on opposite strands of the *E. ruminantium* chromosome and are located on opposite sides of the origin of replication, so recombination between these genes could lead to inversion of the region bounded by the genes. This observation led us to search for such inversions in 12 different *E. ruminantium* isolates, but amplification across the *rho* (Figure 4.2) and *tuf* repeat units with combinations of primers located on either side of the repeat units indicated that the chromosomal arrangement is the same in all the isolates tested. Therefore, although such chromosomal rearrangements may well occur in *E. ruminantium*, the recombinant progeny may not be viable. In fact, although large chromosomal rearrangements are less common within than between species (Hughes 2000a) a high frequency of rearranged genomes has been found in clinical isolates of other pathogenic bacteria such as *Salmonella enterica* serovar Typhi, *Neisseria* spp, *Pseudomonas aeruginosa* and *Bordetella pertussis* (reviewed in Hughes, 2000a). These rearrangements may be favoured as a result of conferring some survival advantage, such as improving the ability of the populations containing them to evade the immune system, however this situation does not seem to have occurred in the case of *E. ruminantium*.

Since chromosomal inversions and translocations are more common between closely related species we determined whether such events have occurred in *Ehrlichia* species that are closely related to *E. ruminantium*. Whole genome comparison showed that there has been inversion around the *rho* genes between *E. ruminantium* and *E. chaffeensis*, but that the arrangement in *E. canis* is the same as that of *E. ruminantium* (Chapter 3, Figure 3.7). The arrangement around the *tuf* repeat units is the same in *E. ruminantium*, *E. chaffeensis* and *E. canis*. It was also found that the *rho* region was only duplicated in the *Ehrlichia* and *Anaplasma* species; the other Rickettsiales only have one copy of *rho*. Two copies of *tuf* were identified in the *Ehrlichia*, *Anaplasma* and *Wolbachia* species, while only one copy was found in the other organisms investigated.

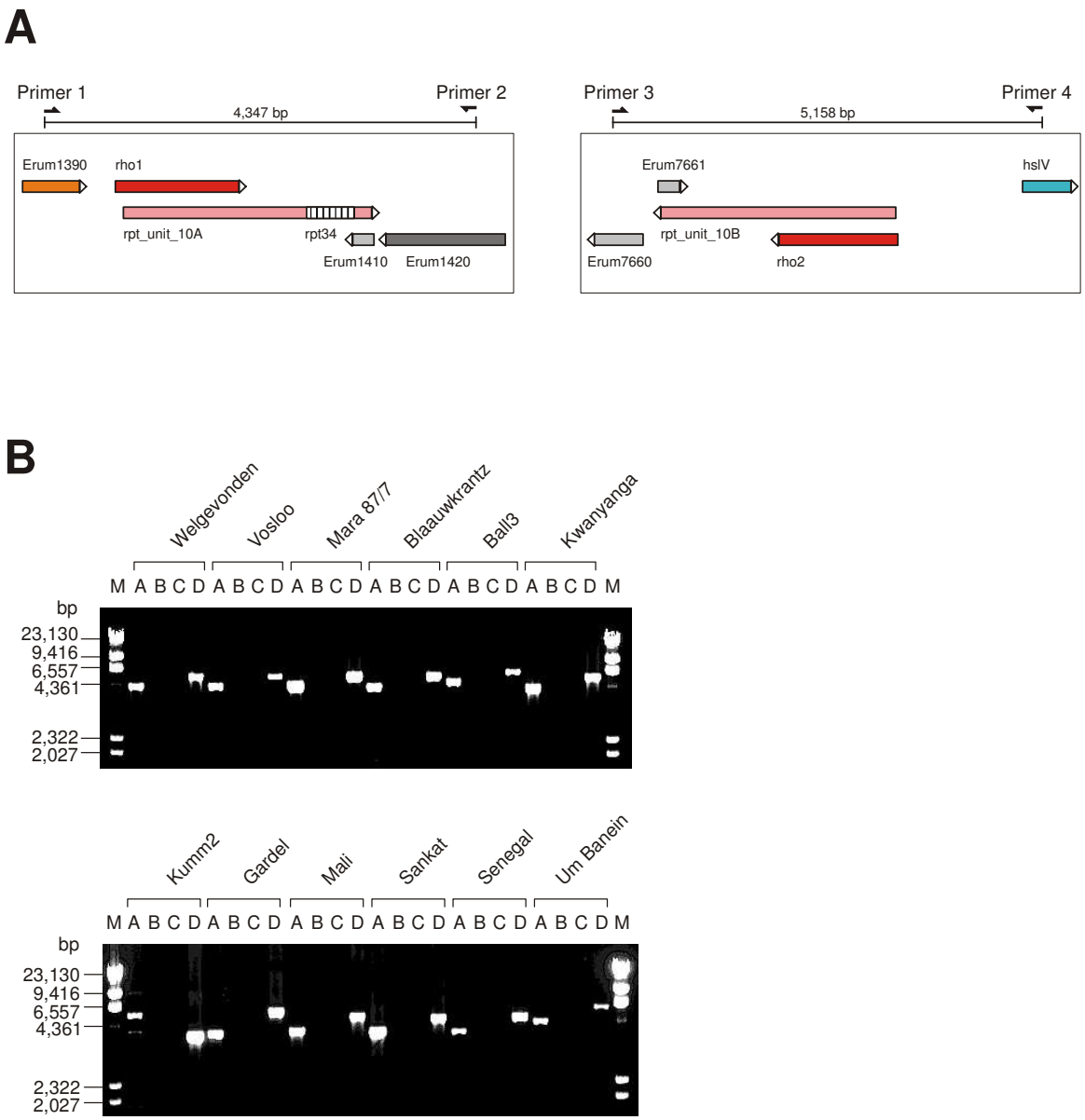


Figure 4.2. PCR amplification across the *rho* repeat regions in *E. ruminantium* isolates.

A. Schematic representation of genes in the two *rho* regions in the *E. ruminantium* (Welgevonden) genome, indicating primer positions. The calculated distance between primers 1 and 2 is 4,347 bp, while primers 3 and 4 are 5,158 bp apart. The inverted repeat units (10A and 10B) are shown in pink while the tandem repeat region (rpt34) is indicated with vertical bars.

B. Gel images of amplicons using the following combinations of primers: lanes A, primers 1 & 2; lanes B, primers 1 & 3; lanes C, primers 2 & 4; lanes D, primers 3 & 4. Lambda *Hin*dIII markers are in lanes labelled M.

4.3.4.2. Duplications appear to generate new genes

In the *E. ruminantium* genome, duplication events appear to have resulted in the formation of several new genes and we will describe four such instances (Figure 4.3 and Figure 4.4).

The first instance concerns repeat units 8A and 8B with 91.3% identity that overlap Erum1100 and the 5' end of Erum1110 (Figure 4.3, panel A). As described in sub-section 4.3.3.1 the 5' ends of these ORFs were 90.8% identical, but Erum1100 does not contain the 27 bp tandem repeat which forms the 3' part of the larger Erum1110 ORF. It appears that either Erum1100 was duplicated and the copy became fused with the tandem repeat to form Erum1100, or that the 5' part of Erum1110 upstream of the tandem repeat was duplicated and the copy became the gene Erum1100.

In the second instance where there appears to have been duplication of a gene we were not able to identify a repeat unit. Two adjacent genes, Erum8170 and Erum8180 show similarity to, respectively, the 3' and 5' ends of the following gene, Erum8190 (Figure 4.3, panel B). It is possible that Erum8190 was duplicated and mutations have arisen such that a stop codon was introduced, splitting the duplicated gene into two. The sequences of Erum8170 and Erum8180 may then have diverged such that their nucleotide sequences now have 56.7% and 60.9% identity respectively to the 3' and 5' ends of the Erum8190 sequence.

In a third example, two genes appear to have been duplicated and fused (Figure 4.3, panel C). The sequence of Erum4120 is 99.1% identical to the 5' end of Erum4140, while the 3' end of Erum4140 has 50.5% identity with the sequence of the following gene Erum4150. It appears likely that Erum4150 was duplicated and the copy mutated until it was 50.5% identical with its parent gene. Subsequently Erum4120 was duplicated and the copy fused with the mutated copy of Erum4120 to generate the new gene Erum4140. The Pfam domains identified in Erum4120 and Erum4150 were also present in Erum4140. Erum4120 is a conserved hypothetical protein and

contains a probable transcriptional regulator domain (PF02082), while Erum4150 has similarity to cysteine desulfurase.

In a fourth example (Figure 4.4) three paralogous genes were identified, Erum2490, Erum2500 and Erum2510. A direct repeat was identified which has resulted in the apparent duplication of the 3' end of Erum2490, creating a small ORF, Erum2500. The repeat and the small ORF were present in all of the southern African isolates examined but not in three West African isolates, suggesting that it arose through a duplication event in southern Africa, or was deleted in an ancestral West African isolate (Pretorius *et al*., 2010). We compared this region with the other *Ehrlichia* species (Figure 4.4) but could not identify orthologs for any of the three ORFs in *E. chaffeensis* or *E. canis.*

It is interesting to note that orthologs of four of the above mentioned genes (Erum1100, Erum1110, Erum8190 and Erum4140) were identified in the *E. ruminantium* Highway isolate by screening of an expression library with immune serum (Barbet *et al.*, 2001), suggesting that the proteins play a role in immune recognition. In the isolated intracellular environment, intrachromosomal recombination and duplication events may be mechanisms used by *E. ruminantium* to increase its antigenic diversity by modifying gene functions and creating new genes.

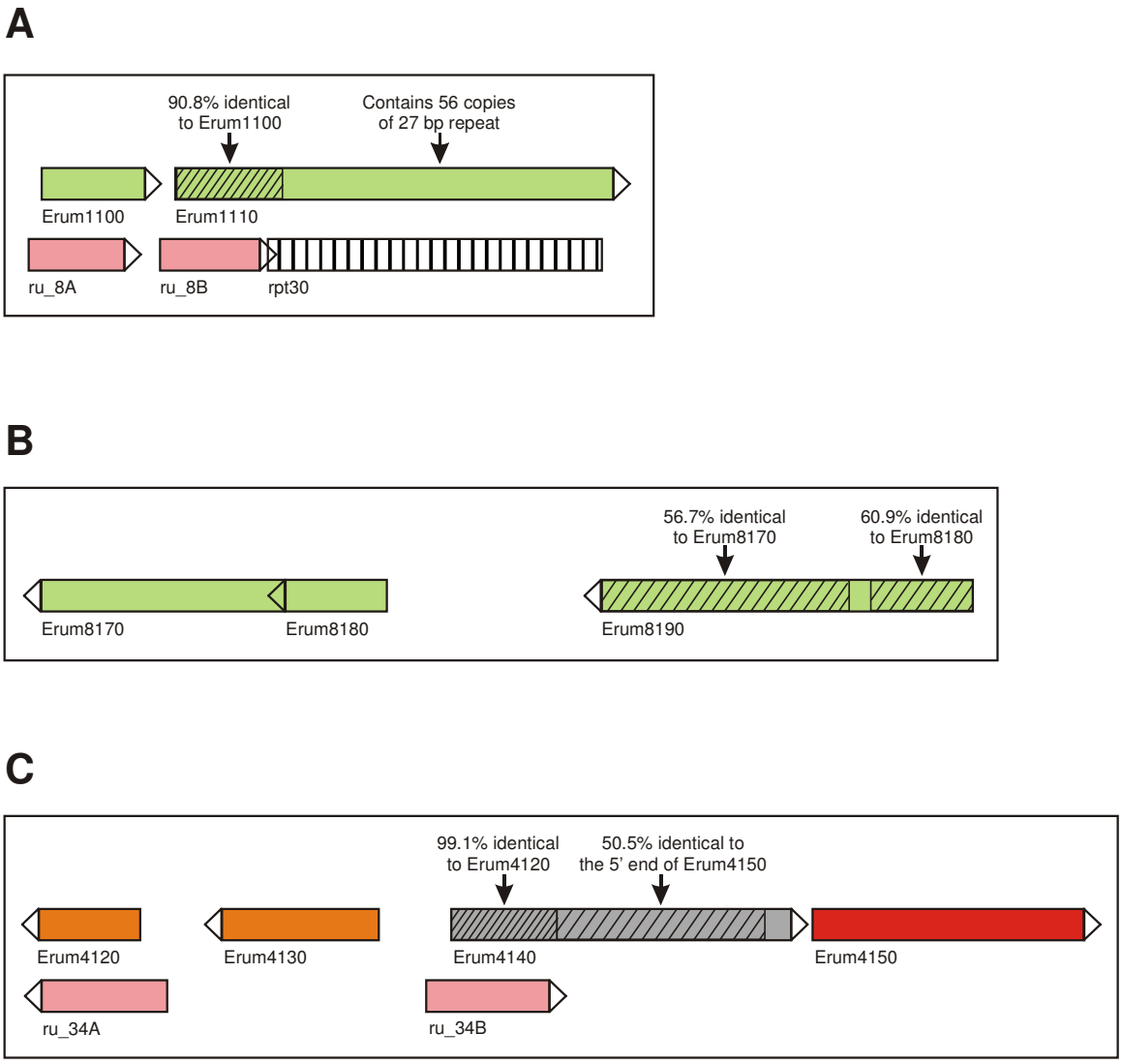


Figure 4.3. Schematic representation of *E. ruminantium* genes that may have arisen through duplication events. Direct and inverted repeat units (ru) are marked in pink and tandem repeats (rpt) are indicated with vertical bars.

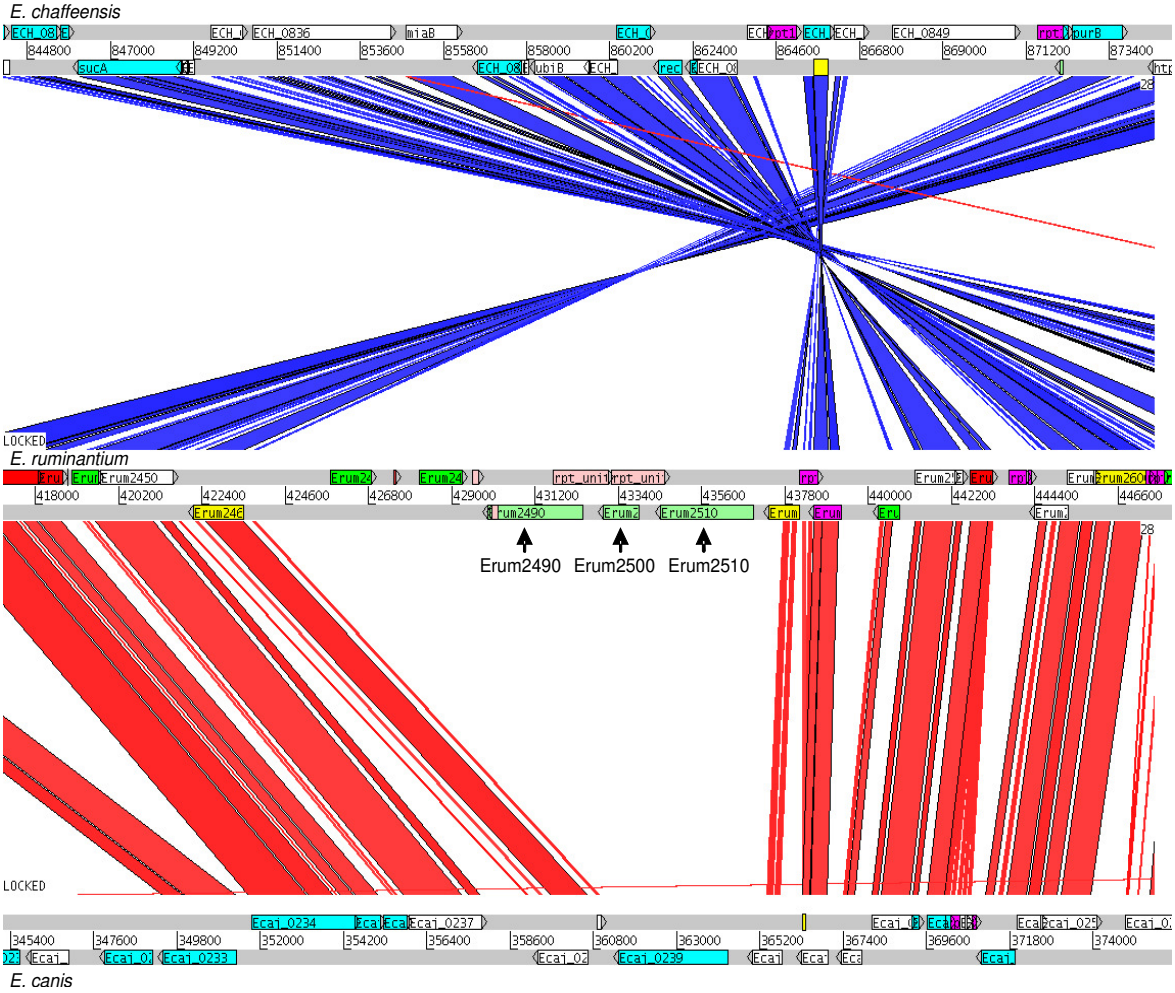


Figure 4.4. Screen capture from ACT of the area around Erum2490, Erum2500 and Erum2510 in *E. ruminantium* (middle), compared to *E. chaffeensis* (top) and *E. canis* (bottom). The grey bars indicate the forward and reverse strands with putative ORFs, while the red and blue lines between the genomes represent the similarities (BLASTn matches) between the three genomes. Direct and inverted repeats are shown in pink.

4.3.5. Ankyrin repeats

Ankyrin repeats are present in a variety of proteins of eukaryotes where they mediate protein-protein interactions. Few examples are found in prokaryotes and the few that exist may originate as a result of horizontal gene transfer from eukaryotic hosts (Bork, 1993). In *E. ruminantium* we identified ankyrin repeat domains in four ORFs: Erum2180, Erum3750, Erum3980 and Erum6220. The functions of all four of these proteins are unknown, although Erum2180 is predicted to code for an 876 aa membrane-associated protein. Erum3750 (5,022 bp encoding 1,674 aa) and Erum3980 (9,006 bp encoding 3,002 aa) are exceptionally large genes, in comparison with the average of 1,032 bp for *E. ruminantium* ORFs, and both genes contain tandem repeats as well. Most of the other Rickettsiales have a small number of genes containing ankyrin repeats, the exceptions are *W. pipientis w*Mel, which contains 23 (Fenn & Blaxter, 2006), *R. felis* with 22 (Ogata *et al*., 2005), and *R. bellii* with 25 (Ogata *et al*., 2006). In *A. phagocytophilum* ankyrin repeats have been implicated in host-pathogen interactions (Caturegli *et al*., 2000), hence these genes may be considered as possible vaccine candidates.

4.4. CONCLUSIONS

Intracellular pathogens have little opportunity for genetic exchange with other bacteria and a process of reductive evolution is predicted to reduce their genetic repertoire (Andersson & Kurland, 1998). This process is thought to occur through intrachromosomal recombination events at repeated sequences which lead to deletions (Rocha, 2003). In the absence of the ability to regain the lost sequences from other bacterial species through horizontal transfer, this process results in the loss of genes whose products must then be obtained from the host. In *E. ruminantium* duplicated and tandemly repeated sequences may be involved in increasing the genetic repertoire of the organism and contribute to the rather larger genome size compared to related organisms. Whatever the role of the repeats, they are maintained and generated in the *E. ruminantium* genome in the face of reductive evolution, suggesting that they provide some selective advantage to the organism.

Table 4.4. Dispersed repeats in the *E. ruminantium* genome. (Adapted from Collins *et al.,* 2005. [Supplementary information])

Identification code	Location of duplication (Co-ordinates)	Length (bp)	% identity	Shortest distance between units (bp)	Feature overlapping repeat unit or within which repeat unit is located
rpt_unit_1A rpt_unit_1B	6938..7180 1465667..1465909	243 243	94.2	57,384	Erum0060, *asd*, aspartate-semialdehyde dehydrogenase Erum8540, truncated aspartate-semialdehyde dehydrogenase
rpt_unit_2A rpt_unit_2B	52203..52297 53409..53503	95 95	96.8	1,112	Erum0400, probable *trmE*, putative tRNA modification GTPase
rpt_unit_3A rpt_unit_3B rpt_unit_3C rpt_unit_3D	60931..61034 61218..61317 61499..61598 63164..63266	104 100 100 103	A-B 91.3 A-C 93.3 A-D 96.2 B-C 96.0 B-D 91.3 C-D 95.1	A-B 184 A-C 465 A-D 2,130 B-C 182 B-D 1,847 C-D 1,566	Erum0440, probable *dksA*, putative DnaK suppressor protein
rpt_unit_4A rpt_unit_4B	129852..130167 130379..130686	316 308	89.6	212	Contains rpt_unit_5A Erum0760, VirB6 fragment. Contains rpt_unit_5B.
rpt_unit_5A rpt_unit_5B rpt_unit_5C	129880..129984 130407..130511 comp(896915..897031)	105 105 117	A-B 97.1 A-C 86.3 B-C 87.2	A-B 423 A-C 749,204 B-C 749,731	Overlaps rpt_unit_4A Erum0760, VirB6 fragment. Overlaps rpt_unit_4B. Erum5240, *virB6*, type IV secretion system protein VirB6
rpt_unit_6A rpt_unit_6B	139431..139499 comp(678172..678240)	69 69	91.3	538,673	 Erum3850, *putA*, proline dehydrogenase/delta-1-pyrroline-5-carboxylate dehydrogenase
rpt_unit_7A rpt_unit_7B	139992..140164 140270..140442	173 173	97.7	106	 Just overlaps the 3' end of Erum0790, *smpB*, SsrA-binding protein
rpt_unit_8A rpt_unit_8B	191251..191697 191850..192296	447 447	91.3	153	Erum1100, unknown Erum1110, unknown
rpt_unit_9A rpt_unit_9B	219584..219673 230376..230465	90 90	100	10,703	Erum1231, probable pseudogene Erum1300, unknown
rpt_unit_10A rpt_unit_10B	241612..244547 comp(1313680..1316410)	2936 2731	91.4	441,557	Erum1400, *rho1*, transcription termination factor 1; Erum1410, unknown Erum7670, *rho2*, transcription termination factor 2; Erum7661, unknown
rpt_unit_11A rpt_unit_11B	253958..254115 254175..254332	158 158	98.1	60	
rpt_unit_12A rpt_unit_12B	283088..284277 comp(1022632..1023821)	1190 1190	100	738,355	Erum1660, *tufA*, elongation factor Tu-A Erum6090, *tufB*, elongation factor Tu-B

Identification code	Location of duplication (Co-ordinates)	Length (bp)	% identity	Shortest distance between units (bp)	Feature overlapping repeat unit or within which repeat unit is located
rpt_unit_13A rpt_unit_13B	341147..341221 767330..767403	75 74	94.7	426,109	 Erum4460, *pccB*, propionyl-CoA carboxylase beta chain
rpt_unit_14A rpt_unit_14B	358642..358823 367881..368062	182 182	90.1	9,058	Erum2090, probable *ftsK*, putative cell division protein FtsK
rpt_unit_15A rpt_unit_15B	429551..429689 comp(430061..430202)	139 142	85.5	372	 Erum2490, unknown
rpt_unit_16A rpt_unit_16B	431691..433107 433209..434603	1417 1395	75	102	Overlaps 5' end of Erum2490, unknown and 3'end of Erum2500, unknown Overlaps 5' end of Erum2500, unknown and 3' end of Erum2510, unknown (cpg1)
rpt_unit_17A rpt_unit_17B rpt_unit_17C	450677..450764 comp(451051..451139) 470611..470698	88 89 88	A-B 89.9 A-C 93.2 B-C 94.4	A-B 287 A-C 19,847 B-C 19,472	Erum2620, conserved hypothetical protein Overlaps rpt_unit18A Overlaps rpt_unit18B
rpt_unit_18A rpt_unit_18B	450729..451763 comp(469987..471020)	1035 1034	98.9	18,224	Overlaps 5' end of Erum2620, conserved hypothetical protein. Contains rpt_unit17B Contains rpt_unit17C
rpt_unit_19A rpt_unit_19B	479859..481027 495479..496644	1169 1166	77.9	14,452	Erum2740, putative integral membrane transport protein Erum2810, putative integral membrane transport protein
rpt_unit_20A rpt_unit_20B	502986..503668 507978..508662	683 685	97.4	4,310	Overlaps 3' end of Erum2840, probable *matA*, putative malonyl-CoA carboxylase Overlaps 5' end of Erum2850, *gatB*, aspartyl/glutamyl-tRNA amidotransferase subunit B Erum2880, truncated malonyl-CoA carboxylase Erum2890, truncated aspartyl/glutamyl-tRNA amidotransferase subunit B
rpt_unit_21A rpt_unit_21B	comp(519062..519456) 525837..526214	395 378	94.7	6,381	Overlaps 3' end of Erum2970, thiC, thiamine biosynthesis protein ThiC Erum3020, truncated thiamine biosynthesis protein ThiC
rpt_unit_22A rpt_unit_22B	526871..527103 839524..839754	233 231	88.5	312,421	Erum4930, unknown
rpt_unit_23A rpt_unit_23B	comp(533409..533615) 534161..534367	207 207	98.1	546	Erum3070, probable nuoC, putative NADH-quinone oxidoreductase chain C Erum3080, truncated NADH-quinone oxidoreductase chain C
rpt_unit_24A rpt_unit_24B	542233..542352 1338617..1338736	120 120	91.7	719,852	Erum3140, putative integral membrane protein
rpt_unit_25A rpt_unit_25B	comp(593577..593499) 597515..597592	79 78	93.7	4,016	Erum3440, *proS*, prolyl-tRNA sythetase

Identification code	Location of duplication (Co-ordinates)	Length (bp)	% identity	Shortest distance between units (bp)	Feature overlapping repeat unit or within which repeat unit is located
rpt_unit_26A rpt_unit_26B	627023..627322 comp(640982..641282)	300 301	99.7	13,660	Erum3601, truncated glutamyl-tRNA(Gln) amidotransferase subunit A Erum3670, *gatA*, glutamyl-tRNA(Gln) amidotransferase subunit A
rpt_unit_27A rpt_unit_27B	642483..642637 757279..757433	155 155	97.4	114,642	Erum4410, putative type IV secretion system protein
rpt_unit_28A rpt_unit_28B	649285..649441 722182..722338	157 157	88.1	72,741	Erum4150, *iscS*, cysteine desulfurase
rpt_unit_29A rpt_unit_29B	709426..709499 954111..954184	74 74	90.5	244,612	Erum4060, *gcp*, O-sialoglycoprotein endopeptidase
rpt_unit_30A rpt_unit_30B	710152..710255 comp(712329.. 712432)	104 104	95.2	2,074	Erum4061, integral membrane protein fragment Erum4070, putative integral membrane protein
rpt_unit_31A rpt_unit_31B	711476..711557 comp(1182628..1182713)	82 86	90.7	471,071	
rpt_unit_32A rpt_unit_32B	713261..713598 comp(751372..751707)	338 336	98.2	37,774	Erum4090, *mdh*, malate dehydrogenase Erum4380, truncated malate dehydrogenase
rpt_unit_33A rpt_unit_33B	comp(717403..717577) 738274..738448	175 175	92.0	20,697	Erum4111, truncated NADH-quinone oxidoreductase chain G Erum4270, *nuoG*, NADH-quinone oxidoreductase chain G
rpt_unit_34A rpt_unit_34B	comp(718146..718707) 719891..720452	562 562	98.8	1,184	Erum4120, conserved hypothetical protein Erum4140, unknown
rpt_unit_35A rpt_unit_35B	comp(730530..730654) 731175..731295	121 121	96.7	521	 Erum4230, putative integral membrane protein
rpt_unit_36A rpt_unit_36B	733191..733281 comp(734025..734115)	91 91	97.8	744	Erum4240, *truA*, tRNA pseudouridine synthase A
rpt_unit_37A rpt_unit_37B	737588..737824 762131..762366	237 236	94.9	24,307	Erum4260, gyrB, DNA gyrase subunit B Erum4431, truncated DNA gyrase subunit B
rpt_unit_38A rpt_unit_38B	740778..741033 comp(742041..742295)	256 255	95.7	1,008	Erum4280, *nuoH*, NADH-quinone oxidoreductase chain H Erum4300, truncated NADH-quinone oxidoreductase chain H
rpt_unit_39A rpt_unit_39B	741012..741170 comp(741562..741720)	159 159	100	392	Erum4280, *nuoH*, NADH-quinone oxidoreductase chain H Erum4290, truncated NADH-quinone oxidoreductase chain H
rpt_unit_40A rpt_unit_40B	741488..741566 1391910..1391988	79 79	89.9	650,344	Just overlaps 3' end of Erum4290, truncated NADH-quinone oxidoreductase chain H
rpt_unit_41A rpt_unit_41B	742755..742937 comp(1191210..1191392)	183 183	96.7	448,273	Erum4310, *gltX2*, glutamyl-tRNA synthetase 2

Identification code	Location of duplication (Co-ordinates)	Length (bp)	% identity	Shortest distance between units (bp)	Feature overlapping repeat unit or within which repeat unit is located
rpt_unit_42A rpt_unit_42B	745635..745759 759352..759474	125 123	96.0	13,593	
rpt_unit_43A rpt_unit_43B	747436..747657 1147611..1147831	222 221	85.0	399,954	Erum4340, unknown. Contains rpt_unit_44A Contains rpt_unit_44C
rpt_unit_44A rpt_unit_44B rpt_unit_44C	comp(747569..747632) 752327..752390 comp(1147743..1147806)	64 64 64	A-B 89.1 A-C 83.8 B-C 92.2	A-B 4,695 A-C 400,111 B-C 395,353	Erum4340, unknown. Overlaps rpt_unit_43A Overlaps rpt_unit_43B
rpt_unit_45A rpt_unit_45B	755346..755441 755853..755948	96 96	95.8	412	Erum4400, unknown Erum4400, unknown
rpt_unit_46A rpt_unit_46B	comp(756803..757070) 770629..770883	268 255	76.3	13,559	Just overlaps 3' end of Erum4400, unknown
rpt_unit_47A rpt_unit_47B	761560..761744 771462..771647	185 186	83.5	9,718	Erum4480, *argB*, acetylglutamate kinase
rpt_unit_48A rpt_unit_48B	comp(763751..764248) 782362..782859	498 498	88.1	18,114	
rpt_unit_49A rpt_unit_49B	765347..765816 800001..800469	470 469	97.9	34,185	
rpt_unit_50A rpt_unit_50B	766583..766746 comp(771245..771408)	164 164	97.0	4,499	Erum4460, *pccB*, propionyl-CoA carboxylase beta chain Erum4471, truncated propionyl-CoA carboxylase beta chain
rpt_unit_51A rpt_unit_51B	comp(795055..795163) 827369..827477	109 109	96.3	32,206	Erum4650, unknown
rpt_unit_52A rpt_unit_52B	comp(826464..826626) 837711..837872	163 162	93.3	11,085	
rpt_unit_53A rpt_unit_53B	842216..842384 865873..866040	169 168	88.8	23,489	Erum4941, truncated dehydrolipoamide dehydrogenase Erum5130, putative dehydrolipoamide dehydrogenase, E3 component of pyruvate or 2-oxoglutarate dehydrogenase complex
rpt_unit_54A rpt_unit_54B	comp(884999..885099) 904220..904319	101 100	90.1	19,121	Erum5210, putative type IV secretion system protein
rpt_unit_55A rpt_unit_55B	906730..906868 907437..907575	139 139	99.3	569	Erum5290, *lipA*, lipoic acid synthetase
rpt_unit_56A rpt_unit_56B	942763..943475 943987..944700	713 714	98.9	512	Overlaps the 3' end of Erum5450, unknown and the 5' end of Erum5460, unknown Overlaps the 3' end of Erum5460, unknown

Identification code	Location of duplication (Co-ordinates)	Length (bp)	% identity	Shortest distance between units (bp)	Feature overlapping repeat unit or within which repeat unit is located
rpt_unit_57A rpt_unit_57B	958202..958558 961162..961512	357 351	83.8	2,604	Contains rpt_unit_58A. Contains rpt_unit_58C. Overlaps 5' end of rpt_unit_59B.
rpt_unit_58A rpt_unit_58B rpt_unit_58C	958372..958519 958576..958723 961328..961473	148 148 146	A-B 87.4 A-C 81.3 B-C 82.7	A-B 57 A-C 2,809 B-C 2,605	Overlaps rpt_unit_57A. Overlaps rpt_unit_59A. Overlaps rpt_unit_57B and rpt_unit_59B.
rpt_unit_59A rpt_unit_59B	958567..959066 961319..961819	500 501	84.9	2,253	Contains rpt_unit_58B. Contains rpt_unit_58C. Overlaps 3' end of rpt_unit_57B.
rpt_unit_60A rpt_unit_60B	982522..982706 comp(983396..983580)	185 185	94.1	690	Erum5681, truncated deaminase Erum5690, putative deaminase
rpt_unit_61A rpt_unit_61B	1030597..1030721 1030805..1030930	125 126	96.8	84	
rpt_unit_62A rpt_unit_62B	comp(1116922..1117320) 1119252..1119650	399 399	99.2	1,932	Erum6610, putative response regulator component of a two-component regulatory system Erum6630, truncated response regulator component of a two-component regulatory system
rpt_unit_63A rpt_unit_63B	1134474..1134664 comp(1135378..1135569)	191 192	95.8	714	
rpt_unit_64A rpt_unit_64B	1165488..1165568 1176413..1176493	81 81	93.8	10,845	Overlaps 3' end of Erum6880, putative integral membrane protein and 5' end of Erum6890, putative integral membrane protein
rpt_unit_65A rpt_unit_65B	1165954..1166233 comp(1303971..1304251)	280 281	92.2	137,738	Erum6890, putative integral membrane protein
rpt_unit_66A rpt_unit_66B	comp(1193676..1193851) 1196553..1196723	176 171	84.8	2,702	Overlaps 3' end of Erum7040, putative cytochrome c oxidase assembly protein
rpt_unit_67A rpt_unit_67B	1229150..1229491 1233153..1233494	342 342	99.4	3,662	Erum7210, truncated uridylate kinase Erum7240, *pyrH*, uridylate kinase
rpt_unit_68A rpt_unit_68B	1235932..1236052 1236097..1236218	121 122	96.7	45	
rpt_unit_69A rpt_unit_69B	1298504..1298724 comp(1299333..1299553)	221 221	100	609	Erum7581, membrane protein fragment Erum7600, putative membrane protein
rpt_unit_70A rpt_unit_70B	comp(1348067..1348189) 1358073..1358195	123 123	93.5	9,884	

Identification code	Location of duplication (Co-ordinates)	Length (bp)	% identity	Shortest distance between units (bp)	Feature overlapping repeat unit or within which repeat unit is located
rpt_unit_71A rpt_unit_71B	comp(1360891..1361021) 1365349..1365479	131 131	90.1	3,105	
rpt_unit_72A rpt_unit_72B	1363980..1364036 1364126..1364182	57 57	91.2	90	
rpt_unit_73A rpt_unit_73B rpt_unit_73C rpt_unit_73D	1381725..1381910 1382357..1382542 1383664..1383849 1385471..1385656	186 186 186 186	A-B 89.8 A-C 90.9 A-D 83.0 B-C 91.9 B-D 84.9 C-D 83.6	A-B 447 A-C 1,754 A-D 3,561 B-C 1,122 B-D 2,929 C-D 1,622	Erum7990, putative integral membrane protein Erum8000, putative integral membrane protein Erum8010, putative integral membrane protein Erum8020, putative integral membrane protein
rpt_unit_74A rpt_unit_74B	comp(1402450..1402705) 1403379..1403633	256 255	96.9	674	Erum8160, *map*, methionine aminopeptidase Erum8161, truncated methionine aminopeptidase
rpt_unit_75A rpt_unit_75B	1463613..1463770 1463871..1464029	158 159	92.6	101	

CHAPTER 5

Selection of possible vaccine candidates

5.1. INTRODUCTION

Vaccines are designed to stimulate a specific protective immune response in humans and animals which are exposed to known specific disease-causing agents and they are considered to be the safest and most cost-effective solution to the control of infectious diseases (Grandi, 2003; Doro *et al*., 2009). Vaccine development comprises the identification of those elements capable of generating immunological protection when administered as a vaccine formulation. Traditionally, this process has involved the isolation, inactivation and injection of the causative microorganism into a susceptible host, followed by extensive biochemical and immunological investigations. For over a century the traditional approach allowed the control and, in some cases, the eradication of many serious infectious diseases such as smallpox and polio (Grandi, 2003). In fact, most commercial vaccines still contain either killed organisms, for example the vaccines against rabies, influenza, plague and cholera, or attenuated microbes, such as the MMR vaccine against measles, mumps and rubella, BCG against tuberculosis and the yellow fever vaccine (http://www.fda/gov/; Grandi, 2003, Serruto & Rappuoli, 2006). Vaccines based on subunits such as toxins detoxified by chemical treatment (diphtheria and tetanus vaccines), purified antigens (hepatitis B and *Bordetella pertussis* vaccines), or polysaccharide conjugated to proteins (meningococcus, pneumococcus and *Haemophilus influenzae* vaccines) are also produced using traditional protocols.

In many instances the traditional methods have failed to generate effective vaccines and yet more modern approaches, such as the development of recombinant subunit or DNA vaccines, have had a limited impact on vaccine production, generating only a few efficacious recombinant vaccines (Grandi, 2003). Examples of commercialised recombinant subunit vaccines include the formulations against *Bordetella pertussis*, hepatitis B virus, *Vibrio cholera*, *Borrelia burgdorferi*

and the human papilloma virus (Kaushik & Sehgal, 2008; http://www.fda/gov/). In recent years vaccine development has been revolutionised by the advances in molecular genetics, DNA sequencing and bioinformatics, and the availability of a growing number of complete microbial genome sequences enables the targeting of possible vaccine candidates starting from genomic information, an approach named reverse vaccinology (Rappuoli, 2000).

The first example of the successful application of reverse vaccinology was the identification of vaccine candidates against serogroup B *Neisseria meningitidis* (Pizza *et al*., 2000). Since then, the approach has been employed in the development of vaccines against several other pathogens, such as *Streptococcus pneumoniae* (Wizemann *et al*., 2001), *S. agalactiae* (Maione *et al*., 2005), *Porphyromonas gingivalis* (Ross *et al*., 2001), *Chlamydia pneumoniae* (Montigiani *et al*., 2002) and *Bacillus anthracis* (Ariel *et al*., 2003). At least two of these vaccines, the *N. meningitides* and *S. agalactiae* formulations, are currently in clinical development (Giuliani *et al*., 2006; Muzzi *et al*., 2007; Serruto *et al*., 2009). Reverse vaccinology has been used to identify putative vaccine candidates for organisms of veterinary importance too, for instance *Dichelobacter nodosus*, the causative agent of ovine footrot (Myers *et al*., 2007), and *Pasteurella multicida* which causes fowl cholera (Al-Hasani *et al*., 2007).

In this chapter the identification of potential vaccine candidates against heartwater will be addressed. Bioinformatic tools were used to select vaccine candidate genes from the genome sequence of *E. ruminantium* (Welgevonden) (Collins *et al*., 2005). The ORFs were evaluated for their ability to induce recall T-cell responses *in vitro* (for the rationale behind this see sub-sections 1.1.5 and 5.3.4) and finally seven ORFs were selected and tested in vaccine formulations for their ability to generate protective immunity in sheep against *E. ruminantium* infection.

5.2. MATERIALS AND METHODS

See Appendix B for materials and media components.

5.2.1. *In silico* selection strategy

The annotation data for each *E. ruminantium* gene, derived as described in Chapters 2-4, were used as the starting point for the selection procedure. ORFs classified into the following categories were considered as possible vaccine candidates: surface-associated or secreted proteins, transporters, proteins putatively involved in the adaptation of bacteria to heat shock and other environmental stresses, proteins of unknown function, proteins containing tetratricopeptide or ankyrin repeats, adhesins, proteases, iron-binding proteins, methyltransferases and GTPases. Homologues of proteins identified as vaccine candidates in other pathogens by means of functional genomics were also included. All ORFs with more than four predicted transmembrane helices and genes tested previously were removed. The remaining ORFs were grouped according to their putative function to facilitate the selection of representatives from each category. The criteria used to decide which genes were selected or rejected are described in more detail in sub-section 5.3.1 and Table 5.2.

5.2.2. Expression of recombinant proteins

5.2.2.1. Directional cloning into the pET vector

Protein expression was performed using the pET102/TOPO® expression system (Invitrogen). Sequence specific primers (Appendix C3) were designed for each of the selected ORFs to facilitate directional cloning into the pET vector. In the case of ORFs having signal peptide coding sequences the 5' primers were designed so as to omit the signal sequences. ORFs larger than 2,000 bp were divided into smaller sub fragments and we also made sure that primer sequences did not overlap large tandem repeat sequences. The ORFs were amplified with *Pfu* polymerase (Promega), a proofreading DNA polymerase that produces blunt ended PCR products. Each 50 µl reaction contained 25 ng *E. ruminantium* (Welgevonden) genomic DNA, 1.25 U *Pfu*

polymerase, 0.2 µM of each primer, 0.2 mM dNTPs, and 1x reaction buffer (containing 2 mM Mg^{2+}). Amplification was carried out on a GeneAmp® PCR System 9700 (Perkin-Elmer Applied Biosystems) under the following conditions: one cycle at 95°C for 2 min, followed by 35 cycles of denaturation (95°C for 30 s), annealing (50°C for 30 s), and extension (72°C for 3 min), with a final extension at 72°C for 7 min. The amplicons were purified with the MSB® Spin PCRapace kit (Invitek) and cloned into the TOPO® pET vector following the manufacturer's protocols. The plasmid constructs were transformed into TOP10 competent *E. coli* cells by electroporation using the Gene pulser™ II (Bio-Rad) as described in the manufacturer's manual. The cells were plated on LB agar plates containing 50 µg/ml ampicillin and incubated overnight at 37°C. Recombinant clones were picked and grown overnight at 37°C in LB broth containing ampicillin (50 µg/ml). The plasmid DNA was purified using the Invisorb® Spin Plasmid Mini *Two* kit (Invitek) and inserts were detected by PCR followed by 1% agarose gel electrophoresis of the amplicons. The reaction mix contained 0.5 µl of plasmid DNA, 0.13 U TaKaRa Ex Taq™ (TaKaRa Bio Inc.), 0.2 mM dNTPs, and 0.25 µM of each of the pET vector specific primers, TrxFus forward and T7 reverse (Appendix C4). The reaction conditions were: one cycle at 94°C for 5 min, 35 cycles of 94°C for 30 s, 50°C for 30 s and 72°C for 3 min, and a final extension at 72°C for 7 min. Clones containing inserts of the correct size were sequenced, using the TrxFus forward and T7 reverse primers, to verify the orientation and sequences of inserts and to ensure that the His-tag was in-frame.

5.2.2.2. Expression and purification of recombinant proteins

We expressed the recombinant proteins using the Overnight Express™ Autoinduction system 1 (Novagen). Aliquots of 100 ml of LB broth containing ampicillin (50 µg/ml) and the Overnight Express™ solutions were inoculated with freshly transformed BL21Star™ (DE3) *E. coli*. The cells were grown overnight at 37°C with shaking and harvested by centrifugation at 3,000 *g* for 10 min at 4°C. The recombinant proteins were extracted from the cell pellets using BugBuster® Protein Extraction Reagent (Novagen) and purified using Protino® Ni 1000 prepacked columns

(Macherey-Nagel) following the manufacturer's instructions. The concentrations of the proteins were determined using the RC/DC Protein Assay (Bio-Rad) and 100 µg aliquots were precipitated for immunological assays. Proteins were precipitated with acetone (8:1 v/v) overnight at -20°C, collected by centrifugation at 10,000 *g* for 10 min and washed with 70% ethanol.

5.2.2.3. Western blot analysis

Expressed proteins were analysed by Anti-His_6 Western blot analysis using standard procedures. The purified proteins were separated on Criterion™ XT precast gels (4-12% gradient, Bio-Rad) at 100 V for approximately 2 h and transferred to PVDF membranes (Millipore Corporation) with a semi-dry blotter (Semi-phor TE70, Hoefer Scientific Instruments) at 110 mA for 90 min. After incubation in blocking buffer (1x PBS, 1% BSA) for 1 h, the blots were incubated overnight at room temperature in the presence of Anti-His_6 antibodies (75 ng/100 ml, Roche) and the following day they were exposed to conjugate [1/20,000 dilution, horseradish peroxide-goat-anti-mouse IgG (Zymed)] for 1 h at room temperature. The membranes were washed three times with wash buffer (1x PBS, 0.05% Tween-20) for 5 min after each incubation step. Finally the recombinant His-tagged protein bands were visualised using SuperSignal®West Pico Chemiluminescent substrate (Pierce) and X-ray film (Roche).

5.2.3. Immunological assays

5.2.3.1. Lymphocyte proliferation assays

Peripheral blood mononuclear cell (PBMC) lymphocyte proliferation assays were performed as described previously (Van Kleef *et al*., 2000; Pretorius *et al*., 2007). Proliferation assays were carried out in triplicate in half-area flat bottomed 96-well plates (Costar) at 37°C in a humidified atmosphere containing 5% CO_2. PBMCs (4 x 10^6/ml) were incubated with the recombinant proteins (1 µg/ml), or partially purified *E. ruminantium* (Welgevonden) antigen isolated from infected bovine endothelial cells (1 µg/well, positive antigen), or uninfected bovine endothelial cell extract (1 µg/well, negative antigen) in a total volume of 100 µl. PBMCs stimulated with

Concanavalin A (ConA) (5 µg/ml, Sigma) were included as a positive control, while wells containing PBMCs without antigen were used as negative controls. The cultures were incubated for 72 h and pulsed with 0.5 µCi/well of [^{3}H] thymidine (Amersham) for the last 6 h of the incubation period. The cells were harvested onto a 96 well glass fibre filter (Wallac) and the [^{3}H] thymidine uptake was determined using a Trilux 1450 Microbeta liquid scintillation and luminescence counter (Wallac).

Results were presented as a stimulation index (SI) averaged from triplicate wells ± standard deviation, where SI was the mean counts per minute (cpm) of immune cells divided by the cpm of naïve cells. *P* value was determined by the one tailed distribution Student's *t*-test. Proliferation assays with a SI ≥ 8 and $P < 0.01$ were considered significant.

5.2.3.2. IFN-γ ELISpot assays

IFN-γ expression was measured by enzyme-linked immunospot (ELISpot) assays in 96-well plates. MAIPS 4510 Multiscreen™-IP filtration plates (Millipore) were coated overnight with mouse anti-bovine IFN-γ mAb CC302 (1 µg/ml) (Celtic Molecular Diagnostics) at 4°C, and washed three times with unsupplemented RPMI-1640. The coated wells were blocked with RPMI-1640 supplemented with 10% FCS for 2 h at 37°C. Freshly isolated PBMCs (4 x 10^6/ml) were added to the wells and stimulated with the recombinant proteins (1 µg/ml) and incubated for 20 h at 37°C in a humidified atmosphere with 5% CO_2. Positive (ConA) and negative (no antigen) controls were included, as already described for the proliferation assays. The plates were washed three times with 0.05% dH_2O-Tween, three times with 0.05% PBS-Tween (PBS-T) and incubated with rabbit anti-bovine IFN-γ anti-serum (Immonodiagnostik) diluted 1/1,500 in PBS-T/1% BSA for 1 h at room temperature. Subsequently the plates were washed four times with 0.05% PBS-T, followed by incubation for 1 h at room temperature with anti-rabbit IgG alkaline phosphatase conjugate (Sigma) diluted 1/2,000 in PBS-T/1% BSA. After six washes with 0.05% PBS-T, 50 µl of substrate solution (Sigma Fast BCIP/NBT substrate tablets) were added and the

plates were incubated in the dark for 15 min. The plates were washed for 2 min under running water and dried overnight. Spot forming cells (SFCs) were counted using an automated ELISpot reader (Zeiss KS ELISPOT Compact 4.5). The number of SFCs produced after stimulation of immune PBMCs with the recombinant proteins was compared to the number of SFCs produced after stimulation of naïve PBMCs with the recombinant proteins. ELISpot samples with 4x the number of spots/million cells compared to the naïve cells were considered positive.

5.2.4. Vaccine trials in sheep

5.2.4.1. Challenge material

Blood stabilate was prepared from an *E. ruminantium* (Welgevonden) infected sheep and titred as reported previously (Brayton *et al*., 2003; Pretorius *et al*., 2007).

5.2.4.2. DNA immunisation

5.2.4.2.1. Cloning of ORFs into pCMViUBs

The ORFs were amplified using specific primers (Appendix C3) containing restriction enzyme sites to facilitate directional cloning into the pCMViUBs vector (Sykes & Johnston, 1999). We searched the sequences of the ORFs for internal restriction sites using the Staden package program Spin (Staden *et al.*, 2000). Of the available recognition sites incorporated in the vector's cloning site, the cutting sites of the endonucleases *Bam*HI and *Sal*I were not present in any of the ORF sequences and were therefore integrated into the primer sequences. PCR amplifications were performed in 100 µl reaction mixtures containing: 50 ng genomic *E. ruminantium* (Welgevonden) DNA, 0.2 mM dNTPs, 0.25 µM of each primer and 0.5 U TaKaRa Ex Taq™ (TaKaRa Bio Inc.) in 1x reaction buffer. The samples were denatured for 2 min at 95°C, followed by 35 cycles of 95°C for 30 s, 50°C for 30 s, and 72°C for 3 min; this was followed by a final extension at 72°C for 10 min. The PCR products were purified with the MSB® Spin PCRapace kit (Invitek) and cloned into the pGEM®-T Easy vector system (Promega) using the protocols provided by the manufacturers. Recombinant cells were grown overnight at 37°C in LB broth

containing 50 µg/ml ampicillin. The plasmid DNA was purified using the Invisorb® Spin Plasmid Mini *Two* kit (Invitek), digested with *Eco*RI (Roche) and inserts were visualised on 1% agarose gels. Clones containing fragments of the expected size were sequenced with the SP6 and T7 primers (Appendix C4). Plasmids containing the correct insert sequence were digested with *Bam*HI and *Sal*I, while the pCMViUBs vector was digested with the same enzymes and dephosphorylated using shrimp alkaline phosphatase (Promega). The ORF inserts and prepared pCMViUBs vector were purified from agarose gels using TaKaRa recochips (TaKaRa Bio Inc.). The inserts were ligated into the linearised dephosphorylated vector using 1 U T4 DNA ligase (Promega). The ligated products were electroporated into TOP10 *E. coli* cells (Invitrogen) using the Gene pulser™ II (Bio-Rad), plated onto LB agar plates containing ampicillin (50 µg/ml) and incubated overnight at 37°C. Positive clones were screened by PCR and sequenced with the vector specific primers, IECO and CMV991 (Appendix C4), to determine whether the correct ORF sequence was present and in-frame.

5.2.4.2.2. Large scale DNA preparation

Cloned ORFs were grown in *E. coli* and the plasmid DNA was purified using NucleoBond® Xtra Maxi purification Kit (Macherey-Nagel) following the manufacturer's instructions. The plasmid DNA was diluted to a final concentration of 1 mg/ml in endotoxin-free PBS (Sigma) and stored at -20°C. An aliquot of each DNA construct was sequenced before being used for immunisation.

5.2.4.2.3. DNA immunisation of sheep

Merino sheep were obtained from a heartwater- and *Amblyomma*-free area and kept in tick-free stables. They were tested for the presence of *E. ruminantium* organisms using the pCS20 real-time PCR assay (Steyn *et al.*, 2008) and divided into groups (Table 5.1). Each animal in groups Experimental 1, Experimental 2 and Negative control 1 received 50 µg plasmid DNA of each ORF construct by intramuscular injection and 5 µg plasmid DNA per ORF precipitated onto gold beads (Biolistic® 1.6 Micron Gold, Bio-Rad) by intradermal gene gun delivery as described previously (Brayton *et al.,* 1997a; Collins *et al.,* 2003; Pretorius *et al.,* 2007). Groups

Experimental 1 and 2 were immunised with a plasmid DNA cocktail containing four and three ORFs respectively, while the Negative control 1 group received empty pCMViUBs vector. Sheep were immunised three times at three week intervals and were needle challenged five weeks after the last immunisation with 10 LD_{50}s of *E. ruminantium* Welgevonden blood stabilate.

The rectal temperatures of the sheep were taken daily from the commencement of the experiment and they were monitored for the onset of clinical symptoms. The severity of the infection was estimated by scoring the clinical signs according to a reaction index (RI) scale (Pretorius *et al.,* 2007). Animals with severe heartwater symptoms were treated with 0.1 ml/kg oxytetracycline (Liquamycin/LA, Pfizer AH) and animals which did not respond were euthanased *in extremis* using 200 mg sodium pentobarbitone (Eutha-Nase, Centaur) per kg body mass.

5.2.4.3. DNA prime–recombinant protein boost immunisation strategy

5.2.4.3.1. Large scale preparation of recombinant proteins

The recombinant proteins were expressed as described in sub-section 5.2.2.2 in 500 ml culture volumes. Two experimental vaccine formulations containing either three or four recombinant proteins were prepared (Table 5.1). The precipitated recombinant proteins were resuspended in endotoxin-free PBS (Sigma) and mixed with adjuvant (Montanide ISA50) (1:1 v/v) on ice using the Ultra Turrax homogenizer (Janke & Kunkel Ika-Labortechnik). The control insert supplied with the TOPO pET kit, the *lacZ* gene, was expressed and used as the negative control recombinant protein (rβ-galactosidase).

5.2.4.3.2. Immunisation of sheep

Sheep that were immunised using the prime–boost strategy (Table 5.1: groups Experimental 3 and Experimental 4) were inoculated twice with the plasmid DNA cocktails as described in sub-section 5.2.4.2.3, followed by 150 µg recombinant protein per ORF by subcutaneous injection, three weeks after the second DNA immunisation. Animals in the negative control group were immunised with the empty pCMViUBs vector followed by 150 µg of recombinant

β-galactosidase protein. The sheep were challenged five weeks after the protein boost and monitored for the onset of clinical symptoms as described in sub-section 5.2.4.2.3.

Table 5.1. The immunisation strategy for the animal trial.

Group	Number of sheep	Inoculated with
Positive challenge control	2	Infected and treated
Negative challenge control	2	None, naïve
Negative control 1	5	3x empty pCMViUBs vector DNA
Negative control 2	5	2x empty pCMViUBs vector DNA, 1x rβ-galactosidase protein
Experimental 1	5	3x ORF cocktail 1* DNA
Experimental 2	5	3x ORF cocktail 2† DNA
Experimental 3	5	2x ORF cocktail 1 DNA, 1x ORF cocktail 1 recombinant protein
Experimental 4	5	2x ORF cocktail 2 DNA, 1x ORF cocktail 2 recombinant protein

* ORF cocktail 1: Erum4470, Erum5430, Erum7300, Erum3630

† ORF cocktail 2: Erum5400, Erum8050, Erum5270

5.3. RESULTS AND DISCUSSION

5.3.1. *In silico* selection of possible vaccine candidates

A reductive strategy was employed to select vaccine candidates from the annotated *E. ruminantium* (Welgevonden) genome sequence. Initially ORFs with functional or structural similarity to proven protective antigens or known virulence factors were identified. From a total of 888 ORFs, 451 were selected and categorised according to their putative functions (Table 5.2, round 1). Since *E. ruminantium* is an obligate intracellular parasite it must be able to invade and survive within host cells and its surface organisation must play a significant part in this process. For this reason surface-associated, membrane-associated and putative exported proteins constituted a large part of the initial selection. Another significant category consisted of proteins of unknown function and many of these, as well as some of the membrane-associated proteins, contained tetratricopeptide or ankyrin repeat domains or tandem repeats. All three repeat elements have been implicated in host-pathogen interactions (Caturegli *et al*., 2000; Core & Perego, 2003; De la Fuente *et al*., 2004; Wilson *et al*., 2005; D'Auria *et al*., 2008; Luo *et al*., 2008; Wakeel *et al*., 2009; Zhang *et al*., 2008a; Zhu *et al*., 2009), hence these genes may be considered as vaccine candidates.

Other possibly important categories include type IV secretion system proteins, transporters and proteases. Proteases have been implicated in pathogenesis (Miyoshi & Shinoda, 2000; Ariel *et al*., 2003, Myers *et al*., 2007) and numerous studies have concluded that type IV secretion systems are essential virulence factors in pathogenic bacteria (Christie, 2001; Lopez *et al.*, 2007; Juhas *et al*., 2008). Other transporters, particularly the ABC transport system, also seem to play an important role in pathogenesis (Basavanna *et al.*, 2009). For example, the iron-binding protein Fbp of *Ehrlichia canis* was found to be immunogenic (Doyle *et al.*, 2005) and *Brucella abortus* Cgt and *Streptococcus pneumoniae* PiuA and PiaA are required for these bacterial pathogens to be fully virulent (Brown *et al*., 2001; Roset *et al*., 2004). PiuA and PiaA are essential for iron uptake too and protect mice against systemic challenge with *S. pneumoniae* (Brown *et al*., 2001).

Furthermore, Pretorius and co-workers have reported that two of the genes included in an *E. ruminantium* experimental DNA vaccine code for ABC transporter ATP-binding proteins (Pretorius *et al*., 2007).

In the second stage of selection the number of candidates was reduced from 451 to 266 (Table 5.2, round 2) by eliminating patented genes (United States Patent 6,593,147; Barbet *et al*., 2001) and ORFs tested previously (Louw *et al*., 2002; Nyika *et al*., 2002; Pretorius *et al*., 2002b, 2007). ORFs with more than four predicted transmembrane helices were also removed from the list for purely practical reasons, since these are often difficult to express (Pizza *et al.*, 2000; Grandi, 2001; Ariel *et al.*, 2003). Practical considerations decreed that we had to reduce the 266 candidates down to a number which could be handled with the resources which were available. In the third round we randomly selected 102 ORFs representing each category (Table 5.2, round 3). The fourth and final round retained most or all of the genes in categories for which there was a more specific functional definition, such as "Type IV secretion system proteins" and "ABC transporters", but made a random selection of representative genes from very broadly defined categories which were well populated, such as "Unknown" and "Membrane-associated". The end result was a manageable final selection of 45 genes.

Table 5.2. Number of ORFs identified as possible vaccine candidates grouped according to their putative function, during several rounds of selection and elimination.

	Round 1	**Round 2**	**Round 3**	**Round 4**
Unknown function	80	65	23	8
Unknown, some miscellaneous information	63	25	16	10
Membrane-associated	149	94	31	5
Exported	25	24	11	3
Type IV secretion system	13	9	4	4
ABC transporters	16	7	4	4
Other transporters	33	12	3	3
Proteases	18	11	3	3
Other*	54	19	7	5
Total	**451**	**266**	**102**	**45**

* Including chaperones, proteins involved in stress responses, and ORFs shown to be protective/ immunogenic in other organisms.

5.3.2. Expression of recombinant proteins

We were able to express 37 of the 45 ORFs identified as possible vaccine candidates. One large ORF was subcloned and expressed as two recombinant proteins giving a total of 38 recombinant proteins. Nine of these were obtained only in a water-soluble form, 14 only as insoluble inclusion bodies, and 15 proteins were obtained as both soluble and insoluble fractions. T-cells recognise proteins in the form of small peptide fragments (Hickling, 1998) and it has previously been shown that recombinant proteins in the form of inclusion bodies could induce cellular immune responses even after denaturation (Leung *et al*., 2004). Hence, insolubility and protein denaturation usually do not affect the outcome of cellular immunological assays. Therefore, all fractions were included in the assays; as a result 53 samples were examined altogether. Figures 5.1 and 5.2 give the *E. ruminantium* identification numbers of the corresponding ORFs and the annotation of these genes can be found in Appendix E.

5.3.3. Physical characteristics of recombinant proteins

In several cases there were differences between the observed and predicted molecular masses of the recombinant proteins. For example, the product of Erum4470 is predicted to be a protein 55.3 kDa in size, whereas the observed molecular mass was 35 kDa smaller at approximately 20 kDa (also see sub-section 5.3.5, Table 5.5 and Figure 5.3). An anomaly in the opposite sense was shown by the recombinant protein encoded by Erum4930, which was 20 kDa larger than its predicted size (results not shown). Some of these discrepancies could result from posttranslational modification or partial protein degradation (Lopez *et al*., 2005), and molecular masses greater than expected have often been attributed to glycosylation. For example, recombinant surface proteins of other rickettsial organisms, specifically *Ehrlichia chaffeensis* P120 and *E. canis* P140 (found to be 33 and 55 kDa larger than predicted) (McBride *et al*., 2000), and MSP1a and MSP1b from *A. marginale* (found to be 27 and 21 kDa larger than expected) (Garcia-Garcia *et al*., 2004) were shown to be glycosylated. Glycosylation appears to be involved in the ability of several Gram-negative bacteria to adhere to and invade host cells (Benz &

Schmidt, 2002), an observation which was corroborated by the adherence of the *A. marginale* MSP1a and the *E. ruminantium* mucin-like protein (Erum1110) to tick cells using an *in vitro* adhesion assay (De la Fuente *et al*., 2003; 2004). Whether any of the larger than predicted *E. ruminantium* proteins used in this study are indeed glycosylated or are involved in adhesion and invasion needs to be elucidated.

5.3.4. Recombinant proteins inducing specific Th1 cellular immune responses

Previous studies have demonstrated that T-cell responses characterised by the expression of IFN-γ are essential in protection against *E. ruminantium* infection (Totté *et al*. 1997; 1999; Mwangi *et al*., 1998; 2002). This was the rationale behind attempting to determine whether any of our *E. ruminantium* recombinant proteins induced proliferation and IFN-γ production *in vitro*. The target lymphocytes were PBMCs from sheep immunised against the parasite by infection and treatment.

A total of 20 recombinant proteins specifically stimulated immune PBMCs to proliferate with SI values ≥ 8, of which 18 were significantly different from the control ($P < 0.01$) (Table 5.3). In addition 17 recombinant proteins elicited an IFN-γ response (>4 spots/million cells) (Figure 5.1, 5.2). Significant lymphocyte proliferation assay responses did not always correspond to positive ELISpot responses. Seven of the recombinant proteins assayed induced both significant PBMC proliferation and IFN-γ production (Figure 5.1, 5.2); they were: Erum3630, Erum4470, Erum5270, Erum5400, Erum5430, Erum7300 and Erum8050. Characteristics of these ORFs are summarised in Table 5.4.

Table 5.3. Lymphocyte proliferation assays using PBMCs from a naïve and an infected and treated sheep stimulated with recombinant proteins. Values in bold indicate significant proliferation (SI ≥ 8, $P < 0.01$).

Antigen	SI_{AVE} Immune	*P* Value
neg Ag	1.3 ± 0.19	0.345
pos Ag	48.7 ± 10.78	0.002
1190i	4.6 ± 1.78	0.032
1960i	5.1 ± 1.14	0.007
2400i	**16.1 ± 2.73**	**0.001**
4840i	**11.0 ± 0.10**	**0.0001**
4860s	**16.4 ± 8.15**	**0.007**
5270s	**12.7 ± 2.23**	**0.001**
5270i	6.6 ± 0.81	0.003
6220s	8.5 ± 2.60	0.018
6220i	7.9 ± 2.55	0.007
6270s	7.7 ± 2.59	0.016
7300s	**8.9 ± 2.13**	**0.003**
7300i	7.6 ± 2.34	0.005
7650i	4.5 ± 2.39	0.040
7780i	**15.4 ± 4.38**	**0.005**
7790i	1.0 ± 0.17	0.013
8050s	7.0 ± 3.31	0.030
8050i	**12.3 ± 2.49**	**0.001**
3790s	6.3 ± 3.45	0.060
3790i	2.4 ± 1.20	0.084
4470s	**13.1 ± 0.74**	**0.001**
4470i	5.5 ± 3.15	0.091
5160s	5.3 ± 2.34	0.029
5160i	4.2 ± 1.39	0.014
5500s	2.4 ± 0.36	0.008
5500i	7.8 ± 0.62	0.00005
5620s	1.7 ± 0.35	0.070
5620i	**12.6 ± 2.69**	**0.006**

Antigen	SI_{AVE} Immune	*P* Value
neg Ag	3.7 ± 0.36	0.002
pos Ag	58.0 ± 7.66	0.0002
5760s	3.4 ± 0.89	0.006
5760i	5.3 ± 0.59	0.0002
7410s	6.9 ± 0.16	0.00008
7410i	3.8 ± 0.32	0.0011
7800s	4.1 ± 2.96	0.012
7800i	3.3 ± 2.99	0.095
0320i	3.8 ± 0.14	0.002
1840i	3.0 ± 0.31	0.002
3110i	2.9 ± 0.88	0.027
4930s	1.9 ± 0.40	0.013
4930i	4.9 ± 1.64	0.009
5430s	**9.2 ± 1.73**	**0.001**
5430i	2.8 ± 0.54	0.005
8270s	4.6 ± 3.40	0.091
0250i	2.3 ± 0.54	0.027
3630s	**35.0 ± 12.29**	**0.005**
2170Bi	3.4 ± 1.03	0.011
3500s	3.1 ± 0.48	0.003
2370s	**11.5 ± 4.14**	**0.004**
2180Ai	**19.7 ± 7.36**	**0.007**
5400s	**15.7 ± 3.04**	**0.001**
2180Bi	**18.9 ± 3.08**	**0.001**
3700s	**9.4 ± 2.95**	**0.005**
1110s	2.0 ± 0.38	0.31
8340s	**14.0 ± 0.02**	**0.0001**
4860i	22.3 ± 10.54	0.026

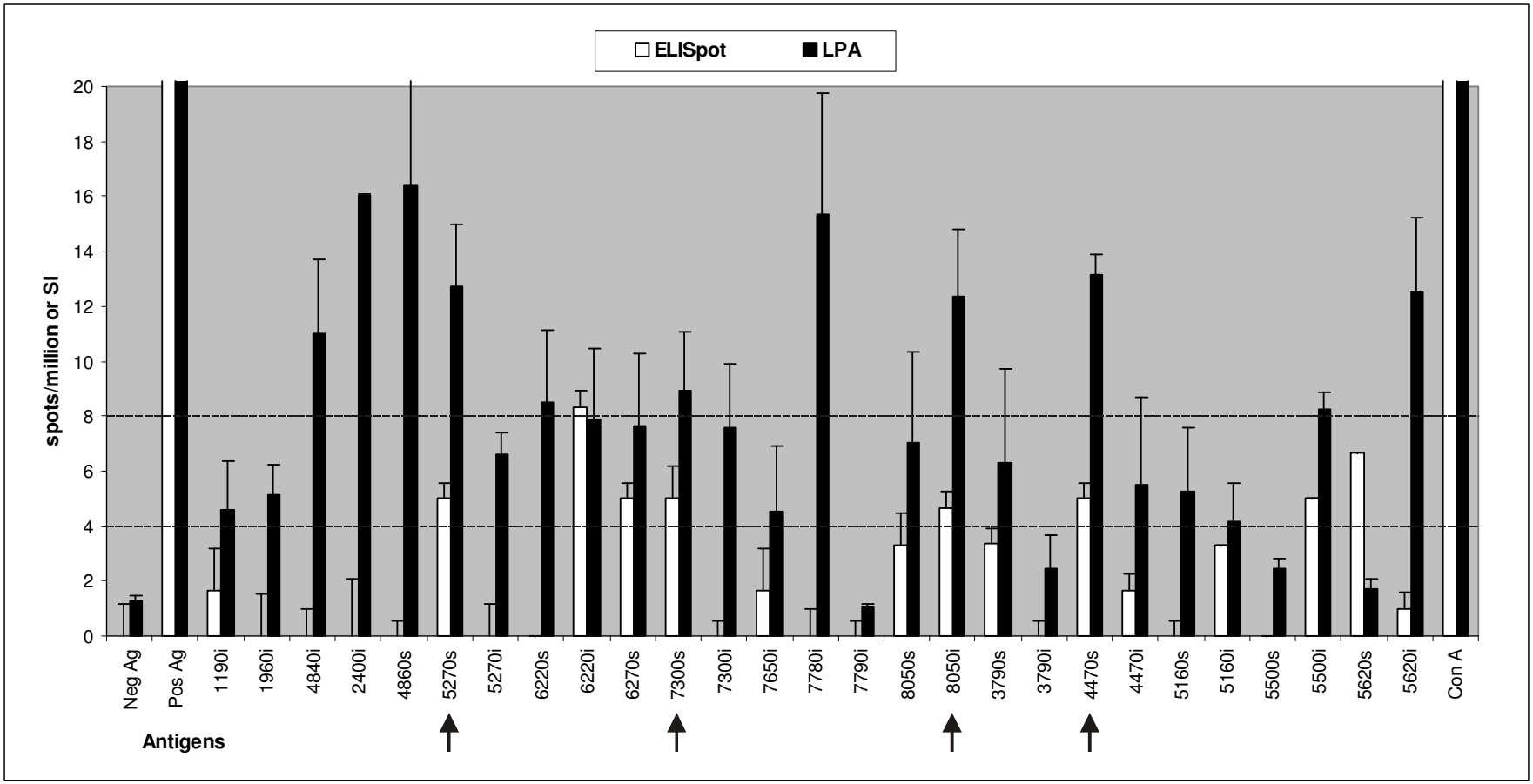


Figure 5.1. ELISpot and lymphocyte proliferation assays (LPA) of PBMCs stimulated with recombinant proteins (plate 1). The **s** and **i** after protein numbers indicate the soluble or insoluble fractions of the proteins. White bars represent the IFN-γ production as spots/million cells, while black bars indicate the SI of the lymphocyte proliferation assays. Samples with more than 4 spots/million cells as well as a SI of more than 8 were selected for animal trials (indicated with arrows).

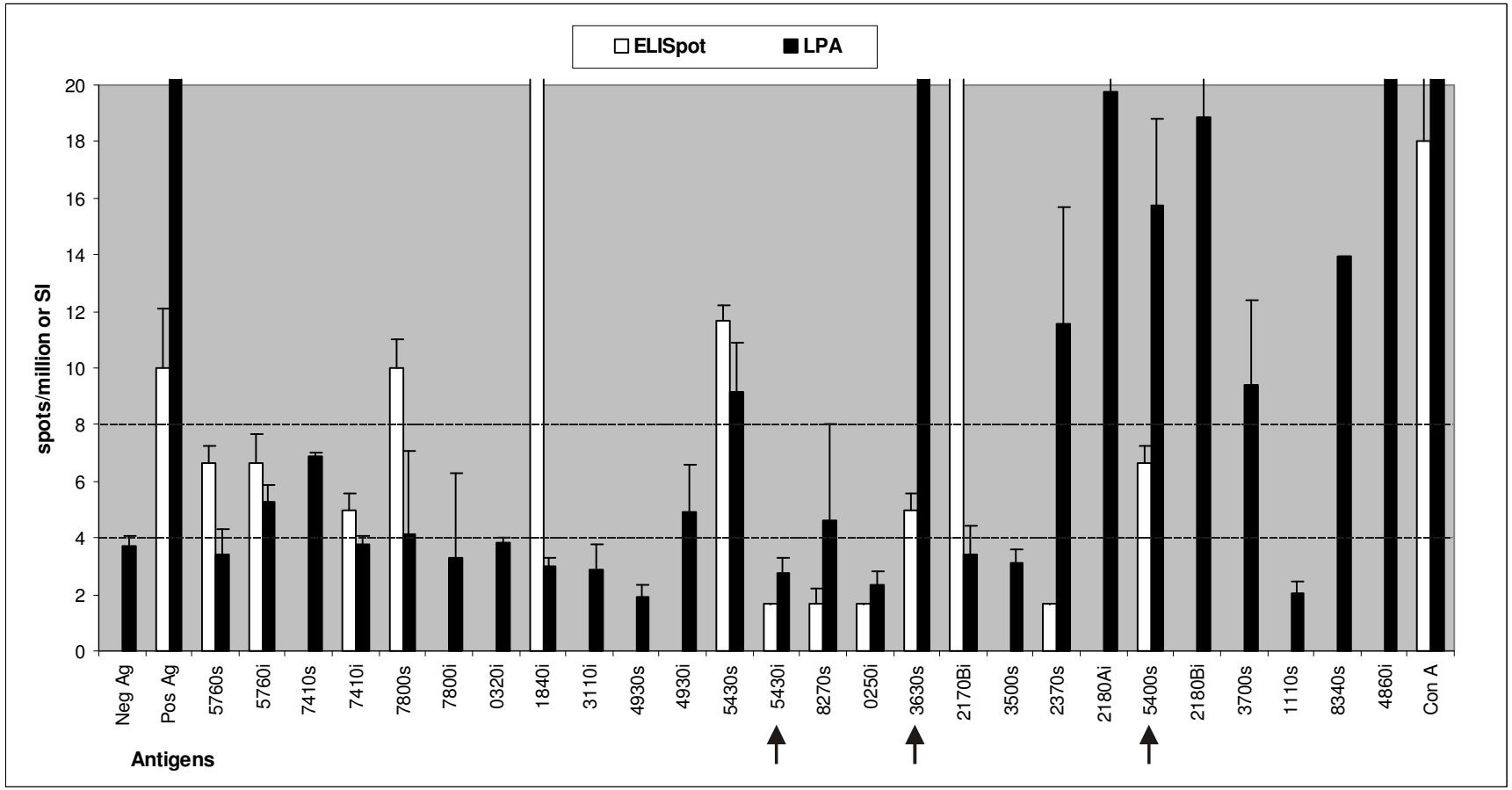


Figure 5.2. ELISpot and lymphocyte proliferation assays (LPA) of PBMCs stimulated with recombinant proteins (plate 2). The **s** and **i** after protein numbers indicate the soluble or insoluble fractions of the proteins. White bars represent the IFN-γ production as spots/million cells, while black bars indicate the SI of the lymphocyte proliferation assays. Samples with more than 4 spots/million cells as well as a SI of more than 8 were selected for animal trials (indicated with arrows).

Table 5.4. Characteristics of the seven ORFs that elicited both significant PBMC proliferation and IFN-γ production *in vitro*. The first column indicates the systematic identification number of each predicted ORF, followed by the gene name (if any), putative protein product and length in number of amino acids. Column 5 shows the transmembrane helices and signal sequences predicted by TMHMM2.0 (Krogh *et al*., 2001) and SignalP3.0 (Nielsen *et al.*, 1997) respectively, while predictions by Phobius (Käll *et al*., 2004) are portrayed in column 6 (th = transmembrane helix). Columns 7 and 8 represent the subcellular localisation predictions by CELLO (Yu *et al*., 2004) and pSORTb2.0 (Gardy *et al*., 2005). Predicted solubility of the expressed proteins as determined using the Recombinant Protein Solubility Prediction algorithm (Harrison, 2000) is indicated in column 9.

Erum ID	Gene name	Protein product	length (aa)	TMHMM & SignalP	Phobius	CELLO	PSORTb	Solubility
3630		membrane protein	519	1 th, signal	signal, 1 th	outer membrane	unknown	34.7%
4470		exported protein	385	signal	signal	outer membrane	outer membrane/multiple	12.9%
5270	*sodB*	superoxide dismutase [Fe]	210	–	–	extra cellular	unknown	54.8%
5400		Unknown	173	–	1 th	outer membrane	unknown	33.0%
5430	*ffh*	signal recognition particle protein	450	–	–	cytoplasmic	cytoplasmic/multiple	20.1%
7300		integral membrane protein	157	2 th	signal, 1 th	extra cellular	unknown	54.6%
8050		exported serine protease	476	signal	signal	outer membrane	periplasmic	9.0%

5.3.5. Vaccine trials in sheep

The protective properties of the seven ORFs encoding the recombinant proteins that induced both significant PBMC proliferation and IFN-γ production were assessed in a vaccine trial. Two vaccination regimens have been used in our laboratory previously, DNA only immunisation and a DNA prime–recombinant protein boost method. A DNA vaccine containing four ORFs, designated 1H12, protected 100% of sheep against a lethal needle challenge in laboratory conditions (Pretorius *et al*., 2007). In another experiment, using the *cpg1* gene, better protection was achieved with the prime–boost system (100%) than with the DNA only immunisation (40%) (Pretorius *et al*., 2010). Therefore, both immunisation regimens were utilised in this study and DNA and protein vaccine formulations containing three or four ORF products were prepared for immunisation.

We cloned the ORFs into the pCMViUBs vector in which they should be expressed as fusion products with ubiquitin, which is designed to enhance CTL responses. Figure 5.3 shows Western blots of seven of the recombinant proteins, and the sizes of only five of them correlated with their predicted sizes (Table 5.5, Figure 5.3). The recombinant proteins of Erum4470 and Erum5400 were much smaller (~20 kDa) than their calculated sizes of 55.3 kDa and 35.8 kDa, respectively. This could be caused by posttranslational modification or partial protein degradation as explained in sub-section 5.3.2. Partial protein degradation may also explain the smaller products, in addition to the products of predicted size, observed for the Erum5270 and Erum7300 recombinant proteins (Figure 5.3).

Table 5.5. Predicted sizes of the seven possible vaccine candidates. Protein molecular weight (MW) was predicted using the program Protein Molecular Weight of the Sequence Manipulation Suite (Stothard, 2000).

ORF	Calculated length of PCR product	Predicted protein MW	Predicted MW plus the Thioredoxin and His-tags	Approximate sizes from Western blots
Erum3630	1488 bp	56.4 kDa	72.4 kDa	65 kDa
Erum4470	1086 bp	39.3 kDa	55.3 kDa	20 kDa
Erum5270	633 bp	24.2 kDa	40.2 kDa	40 kDa
Erum5400	522 bp	19.8 kDa	35.8 kDa	20 kDa
Erum5430	1353 bp	49.6 kDa	65.6 kDa	60 kDa
Erum7300	474 bp	16.4 kDa	32.4 kDa	35 kDa
Erum8050	1365 bp	51.3 kDa	67.3 kDa	60 kDa

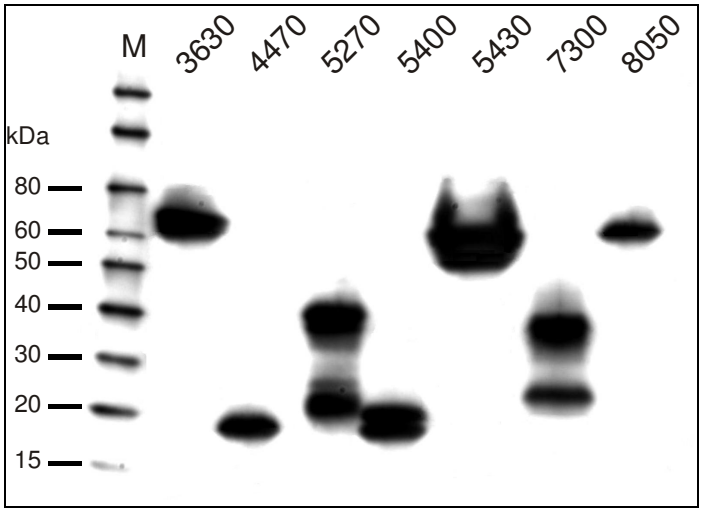


Figure 5.3. Anti-His_6 Western blot of the seven selected ORFs expressed in *E. coli*. Lane M = BenchMark™ His-tagged Protein Standard (Invitrogen)

Five weeks after the final immunisation, the sheep were needle-challenged with a lethal dose of *E. ruminantium* (Welgevonden). All the animals developed severe heartwater symptoms and had to be treated or euthanased, with the exception of one animal (sheep number 6067) in the Experimental 2 group and the infected and treated sheep (Figure 5.4). The animals in group Experimental 2 started to show elevated body temperatures later, as compared to the other groups, their temperatures rose over 40°C only from day 11 onwards (Figure 5.5-8). Temperatures above 40°C were observed for the other experimental groups, and the negative control groups, from day 9. The animals in Experimental 2 were immunised with cocktail 2, which consisted of Erum5270, Erum5400 and Erum8050. The function of Erum5400 is unknown, but the algorithm CELLO predicted that it is located in the outer membrane, though a transmembrane helix was not predicted by the other programs. Erum8050 is predicted to be a serine protease that is exported and Erum5270 codes for iron superoxide dismutase SodB. Superoxide dismutase of *Brucella abortus* elicited protective immunity in mice (Onate *et al*., 2005), while SodB, as part of a multicomponent subunit vaccine or DNA vaccine cocktail, protected against *Mycobacterium avium* infection and induced a Th1 immune response (Park *et al*., 2008; Kathaperumal *et al*., 2009).

We inoculated the animals intramuscularly with 50 µg DNA per ORF, following the protocol which was used for the 1H12 experimental vaccine which had conferred significant protection against lethal needle challenge (Pretorius *et al*., 2007). It is not clear whether one or more of the ORFs contributed to the protection we obtained in the current experiment and if only one of the ORFs was protective it means that the animal received an effective vaccine dose of only 50 µg. Even if all four ORFs induced protection, the animal only received a total dose of 200 µg, which correlates more closely with the doses typically used in mice (10-100 µg), while much larger doses are usually required for larger animals (Doria-Rose & Haigwood, 2003; Dunham, 2002). In fact the dose most often reported for sheep is 500 µg of plasmid DNA per intramuscular inoculation (Chaplin *et al*., 1999; Drew *et al*., 2001; Kennedy *et al*., 2006). It is thus possible that

we could obtain better protection at higher immunisation doses. However, this can only be resolved in a trial where the ORFs are administered individually and at higher doses.

Another aspect to consider is the fact that each vaccine formulation in this experiment contained several constructs. The use of multiple antigens in DNA vaccine formulations can enhance or reduce immune responses. Jiang and co-workers noted a trend of increased T-cell and antibody responses to a pentavalent vaccine cocktail against *Plasmodium falciparum* in comparison to the responses against individual plasmid constructs (Jiang *et al*., 2007). In another study, significant suppression of responses was found when nine plasmid encoding candidate vaccine antigens against *P. falciparum* were pooled (Sedegah *et al*., 2004). We only tested the immune responses of the antigens individually *in vitro*, before the animal trials where they were administered as cocktails. It will therefore be necessary to compare the individual recombinant proteins with the respective cocktails *in vitro* to determine whether there was antigenic interference amongst the antigens.

It has been shown that recombinant protein boosting after primary DNA immunisation can enhance protection against pathogens such as *Mycobacterium tuberculosis* (Wang *et al*., 2004a) and *Leishmania infantum* (Rafati *et al*., 2006). In experiments using the *E. ruminantium* 1H12 ORFs both the recombinant DNA-only immunisation, as well as the recombinant DNA priming followed by recombinant protein boosting, provided 100% protection in laboratory conditions (Pretorius *et al*., 2007; 2008). However only lymphocytes isolated from animals which received a protein boost showed specific proliferation and increased IFN-γ expression when exposed to the recombinant proteins (Pretorius *et al*., 2008). Others have also found that boosting with recombinant protein improved lymphocyte proliferation and increased IFN-γ production (Wang *et al*., 2004b; Rafati *et al*., 2006). In another experiment, using the *cpg*1 gene (Erum2510), protein boosting improved the protection against *E. ruminantium* challenge (Pretorius *et al*., 2010). In the current study, however, protein boosting did not confer any protection. The one immunised animal which survived without treatment was in the Experimental 2 group, which had received

cocktail 2 by DNA-only immunisation, while no animals survived without treatment in the Experimental 4 group, which had also received cocktail 2, in this case via the DNA prime–recombinant protein boost regimen. It is possible that the immunological mechanism responsible for protection is different for individual genes, for instance, it was suggested that *cpg*1 may activate a humoral response (Pretorius *et al*., 2010). From the vaccine development viewpoint this is very disappointing since it complicates the practical experimental issues enormously.

It should be noted that the animals in this experiment were needle challenged. Now there is good evidence that virulent Anaplasmatacea organisms, which are naturally injected by live infected ticks, do not affect the mammalian host in the same way when the organisms are presented as an experimental inoculum in infected blood. One demonstration of this is the well supported finding that animals protected against an *E. ruminantium* needle challenge are not necessarily immune to heartwater-infective ticks (see sub-section 1.1.6.3; Collins *et al*., 2003; Pretorius *et al*., 2008). In another example Galindo *et al*. (2008) showed that immune response genes in sheep infected with *A. phagocytophilum* were differentially expressed in animals experimentally infected as compared to naturally field-infected animals. More importantly, they found that five genes, including IL-2RA, were up-regulated in experimentally infected sheep but down-regulated in naturally tick-infected animals, suggesting that in the latter the adaptive immunity was impaired. Hence a needle challenge does not mimic natural infection and very different results may have been observed in our work if the animals had been challenged with infected ticks. Furthermore, the PBMCs used in the *in vitro* studies were also obtained from experimentally infected sheep. In the future it would be advisable to use heartwater-infective ticks instead of infected sheep blood as the source of virulent *E. ruminantium* organisms, firstly to infect the sheep from which PBMCs are isolated for *in vitro* studies, and then also to challenge the animals used in vaccine trials.

In this study, the reverse vaccinology strategy was not successful in identifying protective antigens against *E. ruminantium*. When using this approach it is crucial to identify the candidates which induce the appropriate immune responses before proceeding to *in vivo* trials, and thus far,

the most effective bacterial vaccine candidates that have been identified are B-cell epitopes of extracellular pathogens (Rappuoli, 2007; Serruto *et al*., 2009). It is generally accepted that the predominant immunological response against obligate intracellular organisms is T-cell mediated, however, detailed knowledge about the immune response against *E. ruminantium* is still lacking. The only cytokine reported to be involved in protection against *E. ruminantium* infection is IFN-γ (Totté *et al*., 1993; 1996) and therefore we used the expression of IFN-γ as one indicator of a relevant immune response. However the seven selected antigens did not protect sheep against a lethal challenge. It is possible that the methods we used to identify IFN-γ production were unreliable, but it is much more likely that IFN-γ expression is not a reliable indicator of a protective immune response against *E. ruminantium* infection. This suggestion is borne out in another recent experiment in our laboratory (Pretorius *et al*., 2008), and other workers also have shown that it is difficult to use IFN-γ expression as a measure of *E. ruminantium* immunity *in vivo* (Vachiéry *et al*., 2006). These observations suggest that we need completely to re-evaluate the role of IFN-γ in protection against heartwater, and this goes some way towards providing plausible reasons for our failure to identify protective *E. ruminantium* genes using the reverse vaccinology approach.

5.4. CONCLUSIONS

Bioinformatic tools were used to identify possible vaccine candidates from the annotated *E. ruminantium* genome sequence. The protective properties of seven ORFs, which induced two different cellular immune responses *in vitro*, were tested in sheep. Only 20% survival was obtained in sheep immunised three times with a DNA formulation consisting of three ORFs; all the other animals succumbed to lethal challenge. The fact that the levels of PBMC proliferation and IFN-γ production did not correlate with each other, nor with the levels of protection, suggests that the current methods being used to select vaccine candidates are just not reliable. In particular it appears that IFN-γ expression alone is not an indicator of protection. We would therefore suggest that other cytokines will have to be included in future immunological studies of the

mechanism of protection against *E. ruminantium* to define in detail what constitutes a protective immune response against this organism. Although reverse vaccinology has been applied successfully in a number of studies the approach was not successful in this study and it still remains a challenge to identify suitable *E. ruminantium* vaccine candidates for further investigation.

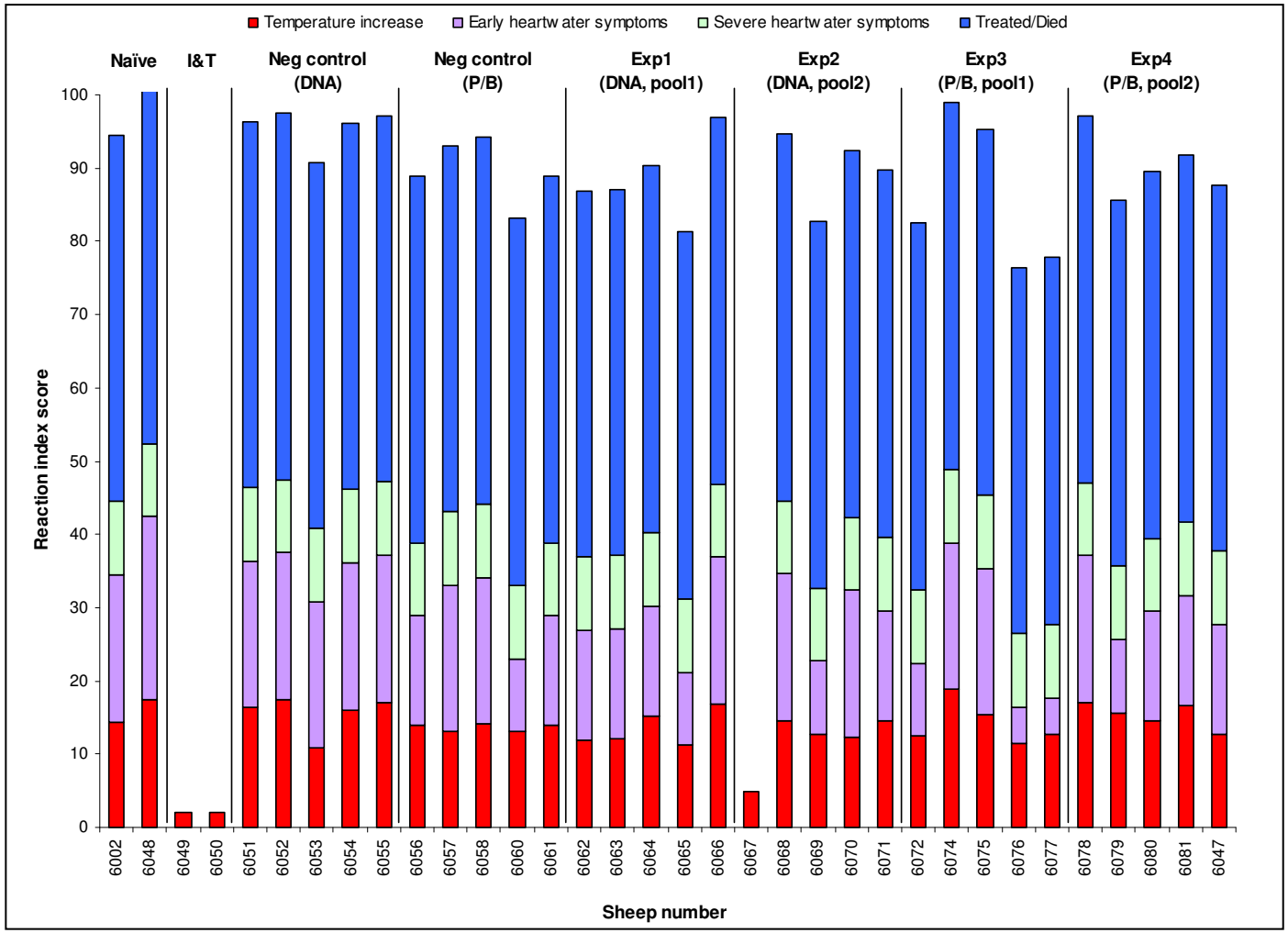
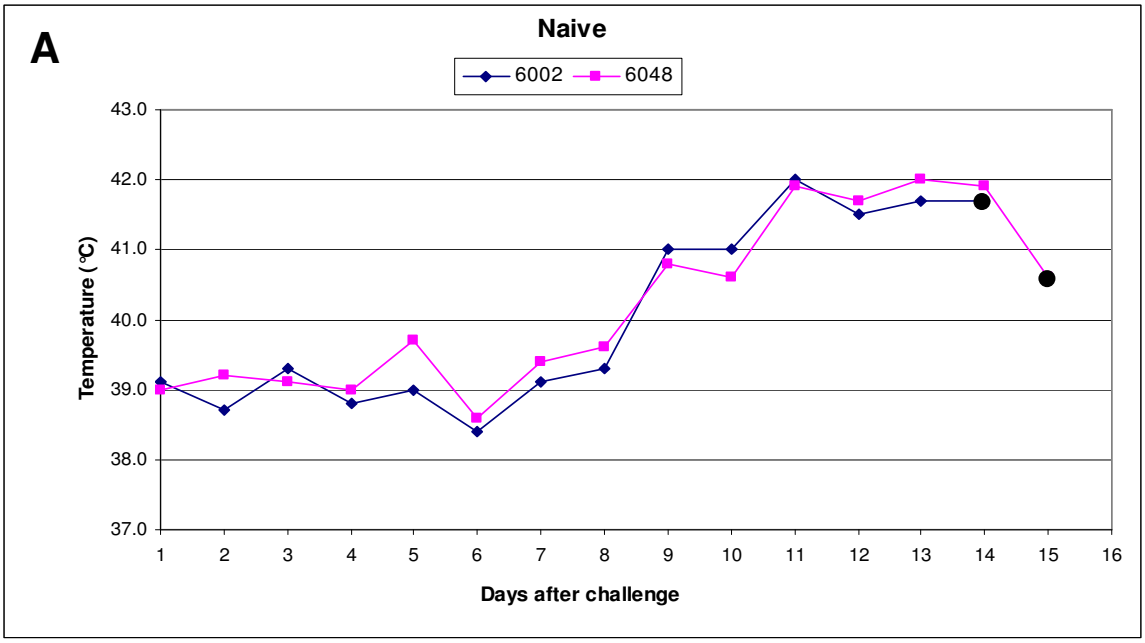


Figure 5.4. Reaction index of sheep. Red blocks represent the total temperature reaction score, while purple indicates early heartwater symptoms, green severe heartwater symptoms, and blue that the animal was treated or euthanased, or died.



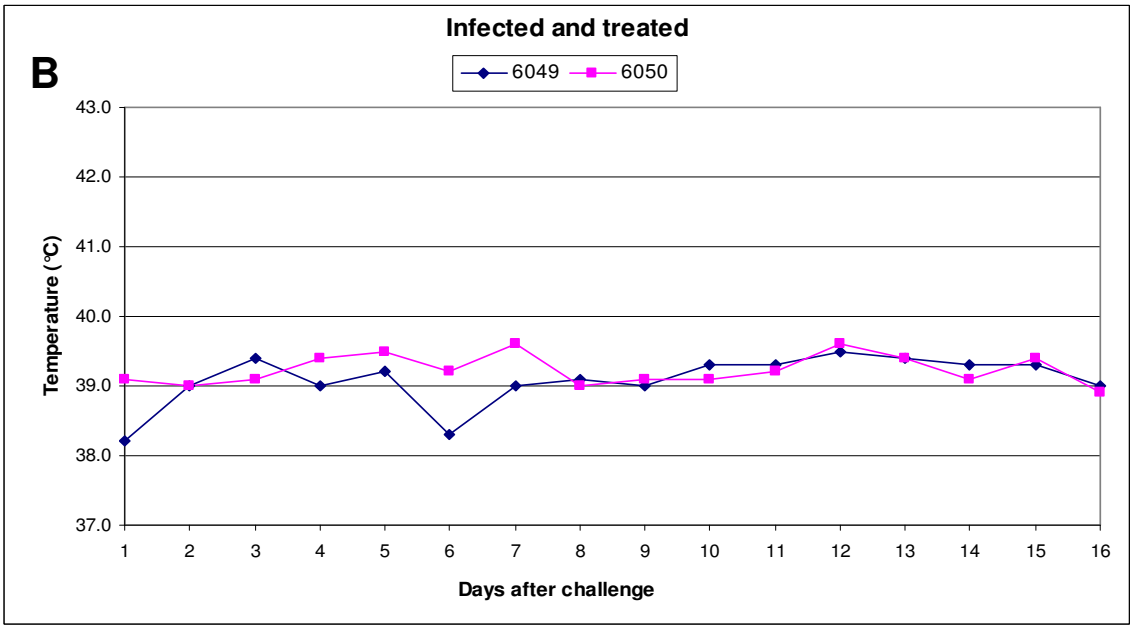
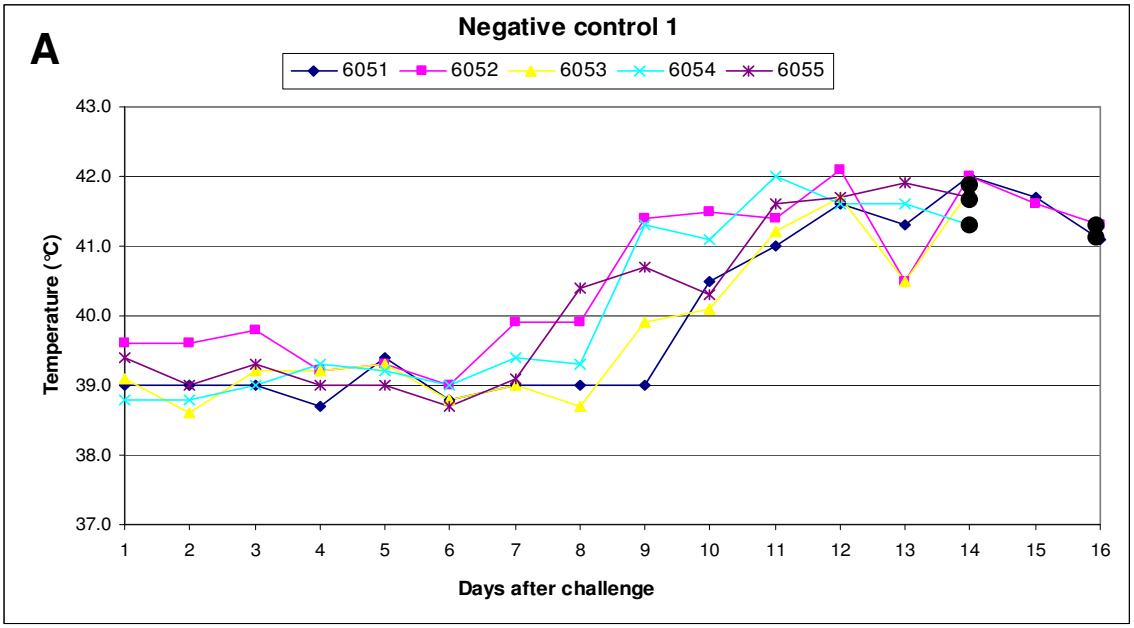


Figure 5.5. Daily post-challenge temperatures of the challenge control group (**A**) and the infected and treated group (**B**). Black dots indicate the day on which the animal died or when it was euthanased. Although both sheep in the infected and treated group survived, temperature measurements are shown only until day 16 after challenge.



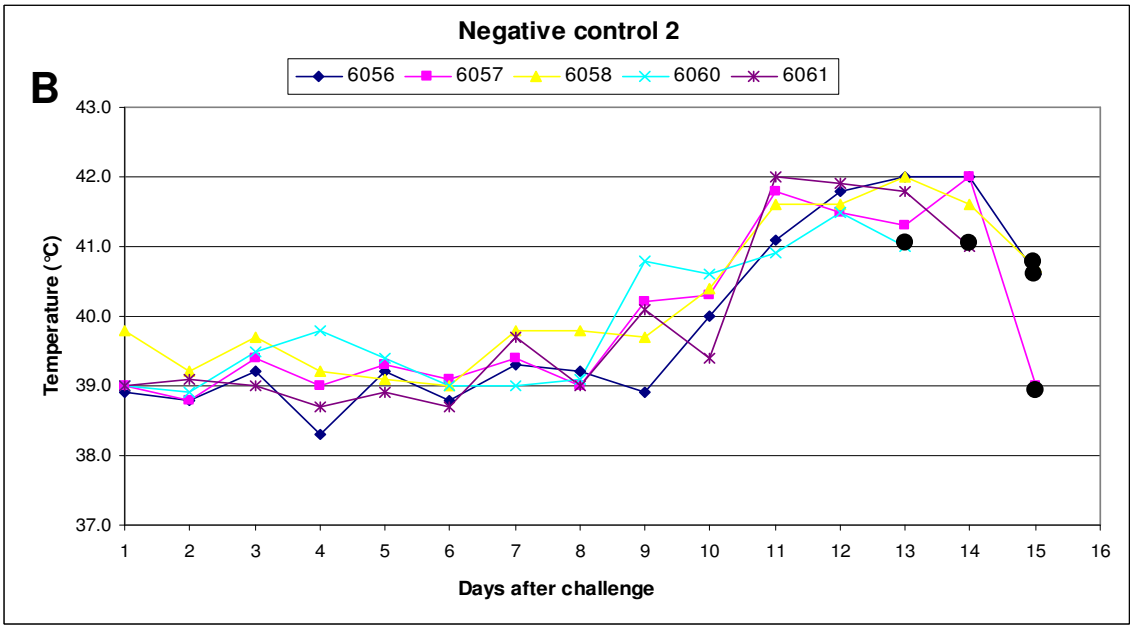
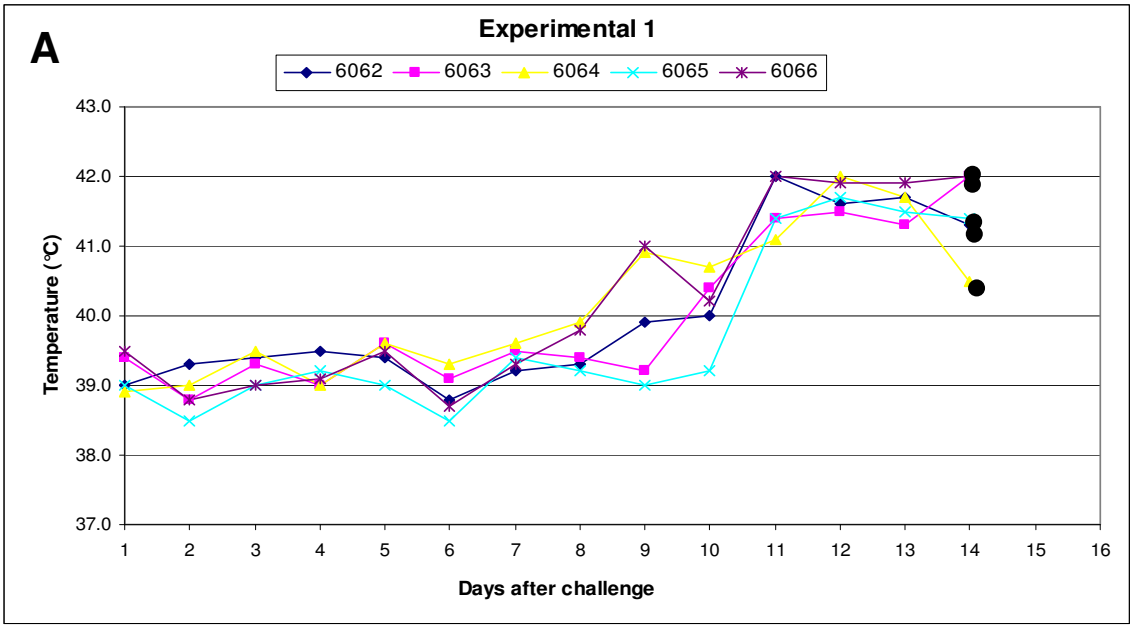


Figure 5.6. Daily post-challenge temperatures of the negative control groups. **A:** Sheep inoculated 3x with empty pCMViUBs vector. **B:** Sheep inoculated twice with empty pCMViUBs vector, followed by a recombinant β-galactosidase protein boost. Black dots indicate the day on which the animal died or when it was euthanased.



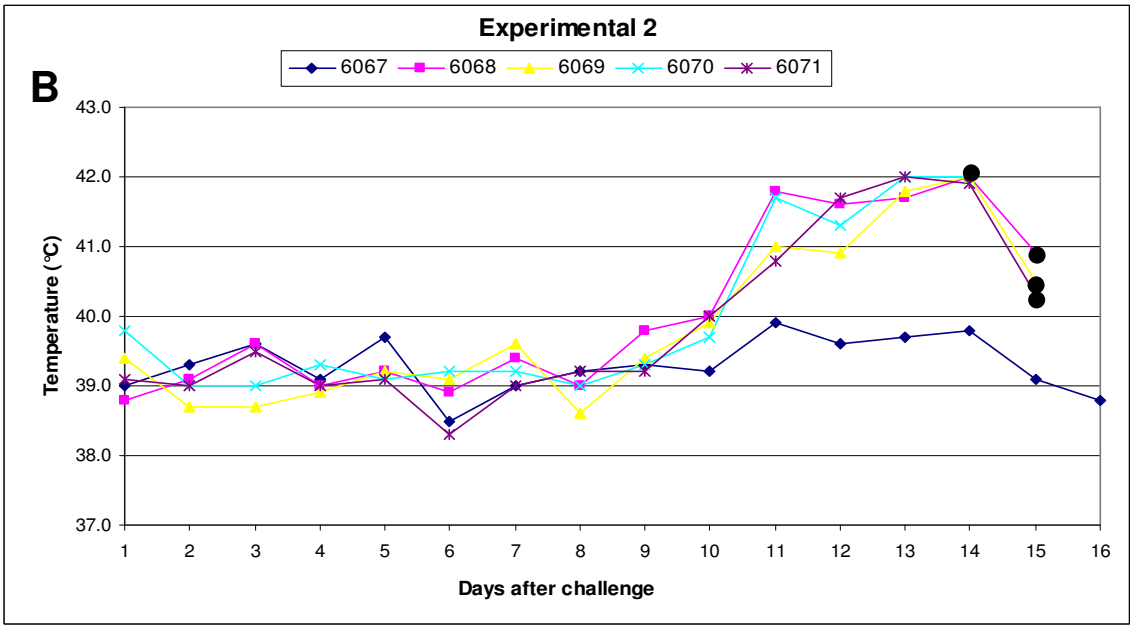
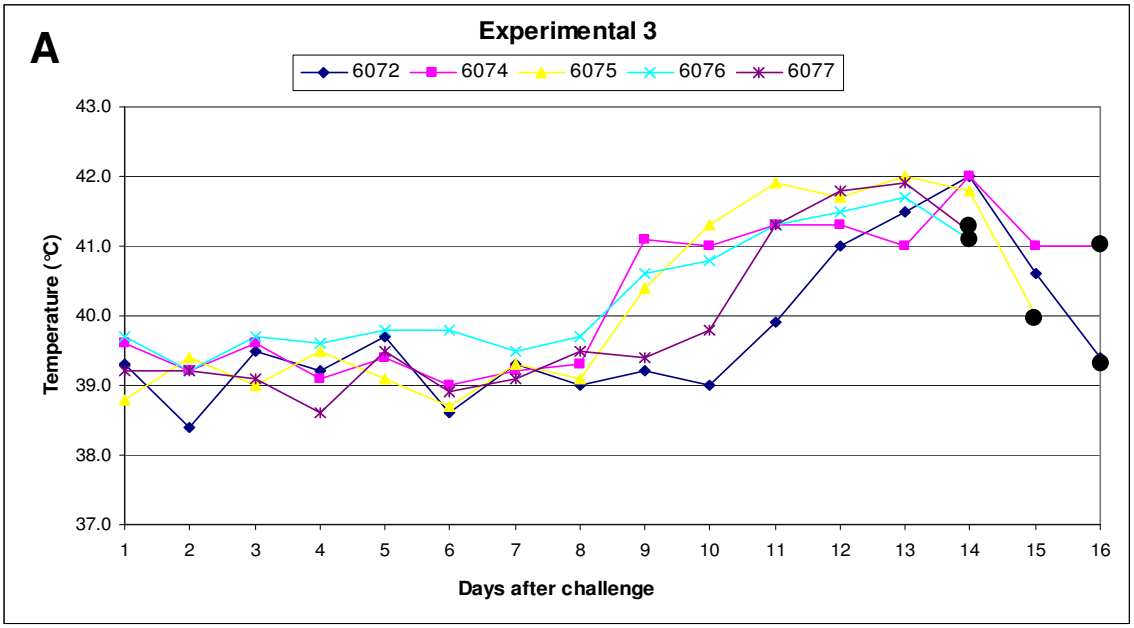


Figure 5.7. Daily post-challenge temperatures of sheep inoculated 3x with ORF cocktail 1 (**A**) or ORF cocktail 2 (**B**) DNA. Black dots indicate the day on which the animal died or when it was euthanased. Temperature measurements of the sheep that survived are shown only until day 16.



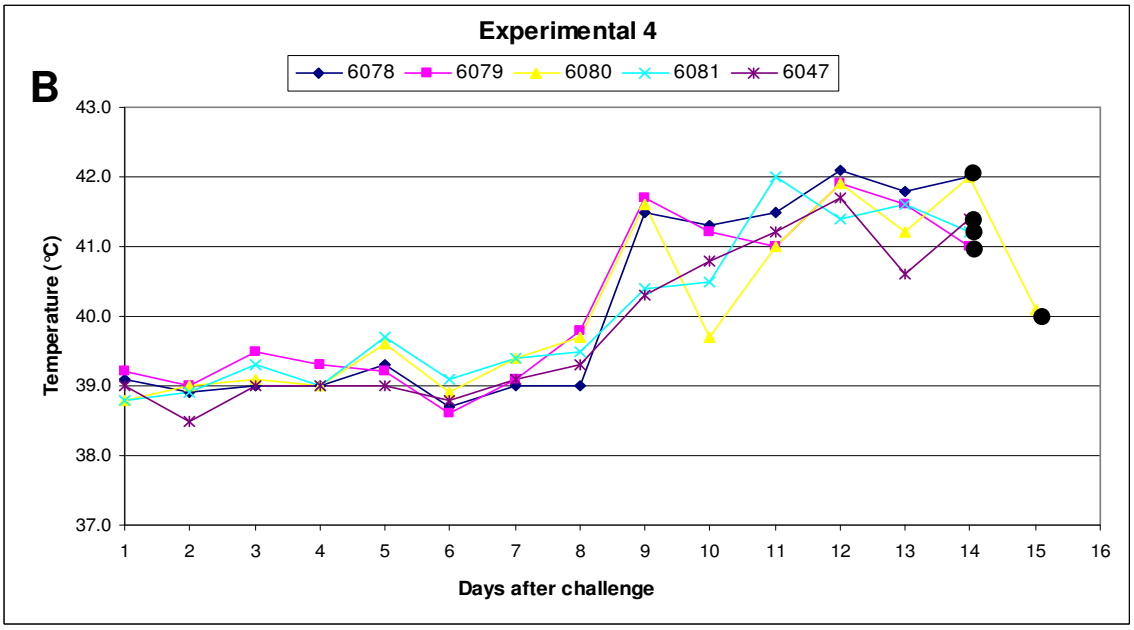


Figure 5.8. Daily post-challenge temperatures of the prime–boost vaccinated groups. **A:** Sheep immunised twice with ORF cocktail 1 DNA followed by an ORF cocktail 1 recombinant protein boost. **B:** Sheep immunised twice with ORF cocktail 2 DNA followed by an ORF cocktail 2 recombinant protein boost. Black dots indicate the day on which the animal died or when it was euthanased.

CHAPTER 6

Concluding discussion

In this thesis the finishing, annotation and analysis of the complete genome sequence of the Welgevonden strain of *E. ruminantium* has been described. The metabolic pathways were constructed, the repetitive sequences of the *E. ruminantium* genome were analysed, and the genome was compared with those of 12 other organisms in the order Rickettsiales. Furthermore, the technique of reverse vaccinology was applied in an attempt to develop an improved recombinant vaccine against heartwater.

Heartwater vaccine development has been hindered by a number of technical difficulties, many of which derive from the fact that obligate intracellular bacteria such as *E. ruminantium* are inherently difficult to study at the molecular genetic level. *E. ruminantium* organisms have exacting culture requirements in eukaryotic cell lines (Zweygarth & Josemans, 2001a; Josemans & Zweygarth, 2002), and are difficult to preserve because of their extreme lability (Oberem & Bezuidenhout, 1987). The isolation of pure *E. ruminantium* DNA, free from host cell DNA contamination, and the construction of representative genomic libraries, have both been shown to be problematic (De Villiers *et al.*, 2000). Because of its intracellular location the genetic manipulation of *E. ruminantium* has not been attempted and therefore little is known about the mechanisms of virulence or pathogenesis. The complete genome sequence can provide us with knowledge of the genetic capabilities of the organism and therefore could provide pointers to ways of surmounting many of the problems noted above. We must note, however, that there are two fundamental difficulties for heartwater vaccine research which can only be addressed directly: there is no reliable small animal disease model (Collins *et al.*, 2003), and there is no satisfactory laboratory-based challenge model (Pretorius *et al.*, 2008). This means that realistic vaccine trials can only be conducted in ruminants which should subsequently be exposed to challenge using infected ticks.

Before the completion of the genome sequence few *E. ruminantium* genes had been characterised; only six genes were located on the published physical and genetic map (De Villiers *et al*., 2000). In fact, most *in vitro* studies of *Ehrlichia* spp. focussed initially on the orthologous immunodominant multigene families discussed in Chapter 2, namely the *E. ruminantium map*1 family (Van Heerden *et al.*, 2004a), the *E. canis* p30 multigene family (Ohashi *et al.*, 1998a), and the p28-Omp locus of *E. chaffeensis* (Ohashi *et al.*, 1998b). MAP1 was identified as one of several dominant immunogenic proteins in serological assays (Van Vliet *et al*., 1994) and later Sulsona and co-workers reported that *map*1 was one member of a multigene family (Sulsona *et al*., 1999). Members of the *map*1 family are differentially transcribed *in vitro* in endothelial and tick cell cultures (Van Heerden *et al*., 2004a; Bekker *et al*., 2005) and *in vivo* in tick midguts and salivary glands (Postigo *et al*., 2007). Host cell-specific expression of the P28 and P30 proteins was also observed (Singu *et al*., 2005; 2006; Peddireddi *et al*., 2009). The differential gene transcription and protein expression of these multigene families suggests that they may play a role in the adaptation of the *Ehrlichia* species to the different cellular environments which the organisms occupy during their lifecycles.

When the whole genome sequences of *Ehrlichia* and *Anaplasma* species became available there was a rapid increase in the numbers of genes and gene families receiving detailed attention. Genes of the type IV secretion system attracted particular interest because they are reported to be involved in pathogenesis (see sub-sections 2.3.2.6 and 5.3.1). In support of this are several studies which have shown that genes coding for type IV secretion system proteins are up-regulated during infection. Lin and co-workers reported that the *A. phagocytophilum* ankyrin repeat protein, AnkA, is delivered to the host cytoplasm via a protein structure that includes VirD4 to facilitate infection (Lin *et al*., 2007), and AnkA of *E. chaffeensis* was found to be translocated into the host-cell nucleus (Zhu *et al*., 2009). In *E. chaffeensis* it was shown that four VirB6 paralogs and VirB9 interact with one another in tick cell culture, presumably to form a functional complex involved in type IV secretion (Bao *et al*., 2009).

In Chapters 3 and 4 the comparison of the *E. ruminantium* genome with the genomes of 12 other members of the order Rickettsiales was described and orthologs of several type IV secretion system genes were found. The four *virB6* genes, *virB9*, and *ankA* are all present in *E. ruminantium*, and this constituted the first indication that *E. ruminantium* has a type IV secretion system. Since this study was performed the number of complete genome sequences in the order Rickettsiales has increased to 39, with 14 in the Anaplasmataceae family and 22 sequences in the Rickettsiaceae family (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html, July 2010). We can anticipate that comparative genomic analysis with the larger number of sequences will improve our understanding of the unique and shared features of the Rickettsiales genomes, and will expand our insights into the varied lifestyles of the different species.

In Chapter 4 it was reported that four of the *E. ruminantium* genes coding for type IV secretion system proteins contain tandem repeats, as do numerous other ORFs. Other workers have shown that proteins containing tandem repeats interact with host cells and facilitate pathogen survival (see sub-section 5.3.1). In addition, Luo and colleagues identified major antibody epitopes in surface-exposed tandem repeat regions of an *E. chaffeensis* and an *E. canis* protein and suggested that these epitopes could be utilised as species-specific diagnostic tools (Luo *et al*., 2009). It appears that *E. ruminantium* is unusual for a small intracellular parasite in that 8.5% of the chromosome is composed of repetitive DNA, and in Chapter 4 evidence was discussed suggesting that these repeats fulfil an important function or functions, although exactly what these are is unclear at present.

In Chapter 5 an attempt to identify vaccine candidates using the reverse vaccinology approach was discussed. With this strategy, possible candidates are selected from the genome sequence using bioinformatics, followed by an *in vitro* screening process. The outcome of reverse vaccinology usually relies on the ability to screen for protective immunity using immunological assays and it is often difficult to find good correlation between positive assays and protection

(Rappuoli, 2001). To complicate matters further, it has been shown that some genes are only expressed *in vivo* and never *in vitro* (Camejo *et al.*, 2009) and as a result cannot be tested in *in vitro* assays. Thus far most successful bacterial vaccines have targeted surface exposed or secreted B-cell epitopes of extracellular pathogens (Serruto *et al.*, 2009) for which *in vitro* immunological assays are relatively straightforward. In the case of obligate intracellular organisms it is generally accepted that the predominant immunological response is T-cell mediated, for which *in vitro* assays are much more complex. Moreover, detailed knowledge about many aspects of the immune response against *E. ruminantium* is still lacking and the selection of appropriate assays remains a problem. Currently we are evaluating the ability of numerous vaccine candidate genes to stimulate the production of various cytokines in cells isolated from blood, spleens and lymph nodes of needle and tick challenged animals, in an attempt to characterise a protective immune response against heartwater. These studies may provide a better insight into the most appropriate *in vitro* immunological assays to use to identify vaccine candidates that are likely to confer protective immunity *in vivo*.

Host immune responses to *Anaplasma* infection have been studied by way of expression profiling. For example, it was found that *A. phagocytophilum* infection in sheep modifies host gene expression and immune responses by activating the inflammatory and innate immune pathways and also impairs adaptive immunity (Galindo *et al.*, 2008). Zivkovic and colleagues determined the effect of *A. marginale* infection on gene expression in the salivary glands of *Rhipicephalus microplus* and discovered genes encoding for putative proteins that are probably required by *A. marginale* for infection and multiplication in ticks (Zivkovic *et al.*, 2010). The genome sequences of several vector and host species have also been completed or are in progress. A draft assembly of the tick *Ixodes scapularis*, vector for the Lyme disease spirochete *Borrelia burgdorferi*, is available and sequencing of the *A. marginale* vector, *Rhipicephalus microplus*, is in progress. Also available are the genomes of the bovine (Bovine Genome Sequencing and Analysis Consortium, 2009) and sheep hosts. The combination of pathogen, vector and host sequence data present new prospects to characterise the inherent structural differences that affect host–pathogen

interactions, and to study metabolic and immunologic pathways implicated in resistance to infection and disease pathology (Zarlenga & Gasbarre, 2009). Investigation of the host–pathogen–vector interactions via transcriptome analyses may also bring us closer to dual-action vaccines for the control of both pathogen transmission and tick infestation (De la Fuente *et al.*, 2010).

Most of the transcriptome studies mentioned above have been conducted using micro-array technology and real-time PCR. With the availability of whole genome sequences and advances in high-throughput sequencing it is possible to address the global features of transcriptomes in a single experiment, with a technique called RNA-Seq (Nagalakshmi *et al.*, 2008) (Chapter 1, subsection 1.2.2.3). Gene expression levels can be assessed from the number of sequence reads related to each gene transcript (Wang *et al*., 2009). The expression levels are quantitative over five orders of magnitude and have been found to be highly reproducible (Mortazavi *et al*., 2008). In addition, RNA-Seq can be used reliably to correct gene annotations based on homology, to define non-coding RNAs and to find new transcripts (Wang *et al.*, 2009). The method has been successfully applied to answer biological questions in a number of organisms, including intracellular bacteria (Cossart & Archambaud, 2009; Albrecht *et al*., 2010), and it is likely to be applied to *E. ruminantium* in the near future.

The ultimate purpose of this study was to identify antigens for inclusion in a recombinant heartwater vaccine. Although promising recombinant vaccine results have been obtained, for *E. ruminantium* and other organisms, the levels of protection obtained using live attenuated vaccines has usually not been matched. The attenuated Welgevonden stock of *E. ruminantium* protects both sheep and goats against a lethal needle challenge (Zweygarth *et al*., 2005; 2008), and preliminary results suggest that the attenuated vaccine can also provide protection against a tick challenge (personnel communication, H. C. Steyn). Although attenuated vaccines are effective, concerns still remain about possible reversion to virulence if the vaccine is to be used in

a non-endemic area. In the case of heartwater, however, this is not a serious problem since the greatest need for a vaccine is the huge endemic area in sub-Saharan Africa.

Targeted genetically attenuated organisms might provide a safe and reproducible platform to develop an efficacious whole-cell vaccine against heartwater, although the obligate intracellular environment of the Rickettsiales is an obstacle to their genetic manipulation. The first successful transformation of a member of the Anaplasmataceae was reported for the murine monocytotropic species *E. muris* (Long *et al*., 2005); more recently it has been shown that it is possible to transform *A. phagocytophilum* by random mutagenesis (Felsheim *et al*., 2006), and *A. marginale* with homologous recombination (Felsheim *et al*., 2010). Using homologous recombination, one could target specific genes or genomic regions for the introduction of foreign genes or to create knock-outs, and this may also provide us with the means to determine the function of the large number of uncharacterised *E. ruminantium* genes. The technique also allows one to generate attenuated vaccines through targeted mutagenesis, as was accomplished in an experimental vaccine against malaria (VanBuskirk *et al*., 2009). These authors introduced gene deletions by double-cross-over recombination to minimise the likelihood of genetic reversion. Currently we are involved in an attempt to identify *E. ruminantium* genes critical for infection by comparing gene expression between the virulent and attenuated Welgevonden strains of *E. ruminantium*. Once the identity of these factors is established, it would be possible to explore the concept of a targeted attenuated vaccine as a reproducible alternative to the current uncharacterised attenuated heartwater vaccine.

Finally, whole genome sequencing has become a standard method for studying living organisms and, since the first complete genome of a free-living organism, *Haemophilus influenzae*, was obtained in 1995, the number of genome sequences in public databases has grown exponentially. To date 1,181 complete bacterial sequences are available in GenBank and more than 3,300 are being sequenced (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, as of July 2010). The availability of the *E. ruminantium* genome sequence, the first complete genome of a free-living

organism to be sequenced and annotated in Africa, will greatly facilitate novel approaches to the study of the organism and its interaction with its hosts. The data derived from this study are vital resources in the search for an efficacious, cost-effective and practical vaccine against heartwater.

Appendix A: References

ADAMS, M.D., CELNIKER, S.E., HOLT, R.A., EVANS, C.A., GOCAYNE, J.D., AMANATIDES, P.G., SCHERER, S.E., LI, P.W., HOSKINS, R.A., GALLE, R.F., *et al*. 2000. The genome sequence of *Drosophila melanogaster*. *Science* **287**: 2185-2195.

AKMAN, L., YAMASHITA, A., WATANABE, H., OSHIMA, K., SHIBA, T., HATTORI, M. & AKSOY, S. 2002. Genome sequence of the endocellular obligate symbiont of tsetse flies, *Wigglesworthia glossinidia*. *Nature Genetics* **32**: 402-407.

ALBRECHT, M., SHARMA, C.M., REINHARDT, R., VOGEL, J. & RUDEL, T. 2010. Deep sequencing-based discovery of the *Chlamydia trachomatis* transcriptome. *Nucleic Acids Research* **38**: 868-877.

AL-HASANI, K., BOYCE, J., MCCARL, V.P., BOTTOMLEY, S., WILKIE, I. & ADLER, B. 2007. Identification of novel immunogens in *Pasteurella multocida*. *Microbial Cell Factories* **6**: 3.

ALLRED, D.R., MCGUIRE, T.C., PALMER, G.H., LEIB, S.R., HARKINS, T.M., MCELWAIN, T.F. & BARBET, A.F. 1990. Molecular basis for surface antigen size polymorphisms and conservation of a neutralization-sensitive epitope in *Anaplasma marginale*. *Proceedings of the National Academy of Sciences of the United States of America* **87**: 3220-3224.

ALLSOPP, M.T.E.P., VISSER, E.S., DU PLESSIS, J.L., VOGEL, S.W. & ALLSOPP, B.A. 1997. Different organisms associated with heartwater as shown by analysis of 16S ribosomal RNA gene sequences. *Veterinary Parasitology* **71**: 283-300.

ALLSOPP, M.T.E.P., THERON, J., COETZEE, M.L., DUNSTERVILLE, M.T. & ALLSOPP, B.A. 1999. The occurrence of *Theileria* and *Cowdria* parasites in African buffalo (*Syncerus caffer*) and their associated *Amblyomma hebraeum* ticks. *Onderstepoort Journal of Veterinary Research* **66**: 245-249.

ALLSOPP, B.A., BEZUIDENHOUT, J.D. & PROZESKY, L. 2004. Heartwater. In: Infectious diseases of livestock 2nd ed. Volume 1. Edited by J.A.W. Coetzer & R.C. Tustin. Oxford: University Press, pp 507-535.

ALTSCHUL, S.F., GISH, W., MILLER, W., MYERS, E.W. & LIPMAN, D.J. 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403-410.

ALTSCHUL, S.F., MADDEN, T.L., SCHAFFER, A.A., ZHANG, J., ZHANG, Z., MILLER, W. & LIPMAN, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389-3402.

ANDERSSON, S.G.E. & KURLAND, C.G. 1998. Reductive evolution of resident genomes. *Trends in Microbiology* **6**: 263-268.

ANDERSSON, S.G.E., ZOMORODIPOUR, A., ANDERSSON, J.O., SICHERITZ-PONTÉN, T., ALSMARK, U.C.M., PODOWSKI, R.M., NÄSLUND, A.K., ERIKSSON, A.S., WINKLER, H.H. & KURLAND, C.G. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* **396**: 133-140.

ANSORGE, W.J. 2009. Next-generation DNA sequencing techniques. *New Biotechnology* **25**: 195-203.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**: 796-815.

ARIEL, N., ZVI, A., MAKAROVA, K.S., CHITLARU, T., ELHANANY, E., VELAN, B., COHEN, S., FRIEDLANDER, A.M. & SHAFFERMAN, A. 2003. Genome-based bioinformatic selection of chromosomal *Bacillus anthracis* putative vaccine candidates coupled with proteomic identification of surface-associated antigens. *Infection and Immunity* **71**: 4563-4579.

BANKIER, A.T., WESTON, K.M. & BARRELL, B.G. 1987. Random cloning and sequencing by the M13/dideoxynucleotide chain termination method. *Methods in Enzymology* **155**: 51-93.

BAO, W., KUMAGAI, Y., NIU, H., YAMAGUCHI, M., MIURA, K. & RIKIHISA, Y. 2009. Four VirB6 paralogs and VirB9 are expressed and interact in *Ehrlichia chaffeensis*-containing vacuoles. *Journal of Bacteriology* **191**: 278-286.

BARBET, A.F., WHITMIRE, W.M., KAMPER, S.M., SIMBI, B.H., GANTA, R.R., MORELAND, A.L., MWANGI, D.M., MCGUIRE, T.C. & MAHAN, S.M. 2001. A subset of *Cowdria ruminantium* genes important for immune recognition and protection. *Gene* **275**: 287-298.

BARRÉ, N., UILENBERG, G., MOREL, P.C. & CAMUS, E. 1987. Danger of introducing heartwater on to the American mainland: potential role of indigenous and exotic *Amblyomma* ticks. *Onderstepoort Journal of Veterinary Research* **54**: 405-417.

BASAVANNA, S., KHANDAVILLI, S., YUSTE, J., COHEN, J.M., HOSIE, A.H., WEBB, A.J., THOMAS, G.H. & BROWN, J.S. 2009. Screening of *Streptococcus pneumoniae* ABC transporter

mutants demonstrates that LivJHMGF, a branched-chain amino acid ABC transporter, is necessary for disease pathogenesis. *Infection and Immunity* **77**: 3412-3423.

BATEMAN, A., COIN, L., DURBIN, R., FINN, R.D., HOLLICH, V., GRIFFITHS-JONES, S., KHANNA, A., MARSHALL, M., MOXON, S., SONNHAMMER, E.L., *et al*. 2004. The Pfam protein families database. *Nucleic Acids Research* **32**: D138-D141.

BEKAL, S., CRAIG, J.P., HUDSON, M.E., NIBLACK, T.L., DOMIER, L.L. & LAMBERT, K.N. 2008. Genomic DNA sequence comparison between two inbred soybean cyst nematode biotypes facilitated by massively parallel 454 micro-bead sequencing. *Molecular Genetics and Genomics* **279**: 535-543.

BEKKER, C.P., POSTIGO, M., TAOUFIK, A., BELL-SAKYI, L., FERRAZ, C., MARTINEZ, D. & JONGEJAN, F. 2005. Transcription analysis of the major antigenic protein 1 multigene family of three in vitro-cultured *Ehrlichia ruminantium* isolates. *Journal of Bacteriology* **187**: 4782-4791.

BELL-SAKYI, L., ZWEYGARTH, E., BLOUIN, E.F., GOULD, E.A. & JONGEJAN, F. 2007. Tick cell lines: tools for tick and tick-borne disease research. *Trends in Parasitology* **23**: 450-457.

BENNETT, S.T., BARNES, C., COX, A., DAVIES, L. & BROWN, C. 2005. Toward the 1,000 dollars human genome. *Pharmacogenomics* **6**: 373-382.

BENSON, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**: 573-580.

BENTLEY, D.R. 2006. Whole-genome re-sequencing. *Current Opinion in Genetics and Development* **16**: 545-552.

BENZ, I. & SCHMIDT, M.A. 2002. Never say never again: protein glycosylation in pathogenic bacteria. *Molecular Microbiology* **45**: 267-276.

BESEMER, J., LOMSADZE, A. & BORODOVSKY, M. 2001. GeneMarkS: a self-training method for predicition of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Research* **29**: 2607-2618.

BEZUIDENHOUT, J.D., PATERSON, C.L. & BARNARD, B.J.H. 1985. *In vitro* cultivation of *Cowdria ruminantium*. *Onderstepoort Journal of Veterinary Research* **52**: 113-120.

BEZUIDENHOUT, J.D. 1987. Natural transmission of heartwater. *Onderstepoort Journal of Veterinary Research* **54**: 349-351.

BONFIELD, J.K., SMITH, K. & STADEN, R. 1995. A new DNA sequence assembly program. *Nucleic Acids Research* **23**: 4992-4999.

BORK, P. 1993. Hundreds of ankyrin-like repeats in functionally diverse proteins: mobile modules that cross phyla horizontally? *Proteins* **17**: 363-374.

Bovine Genome Sequencing and Analysis Consortium. 2009. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**: 522-528.

BRASLAVSKY, I., HEBERT, B., KARTALOV, E. & QUAKE, S.R. 2003. Sequence information can be obtained from single DNA molecules. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 3960-3964.

BRAYTON, K.A., BOTHMA, G.C., VOGEL, S.W. & ALLSOPP, B.A. 1997a. Development of the OPgun™ for bombardment of animal tissues. *Onderstepoort Journal of Veterinary Research* **64**: 153-156.

BRAYTON, K.A., FEHRSEN, J., DE VILLIERS, E.P., VAN KLEEF, M. & ALLSOPP, B.A. 1997b. Construction and initial analysis of a representative λZAPII expression library of the intracellular rickettsia *Cowdria ruminantium*: cloning of *map*1 and three other *Cowdria* genes. *Veterinary Parasitology* **72**: 185-199.

BRAYTON, K.A., DE VILLIERS, E.P., FEHRSEN, J., NXOMANI, C., COLLINS, N.E. & ALLSOPP, B.A. 1999. *Cowdria ruminantium* DNA is unstable in a SuperCos1 library. *Onderstepoort Journal of Veterinary Research* **66**: 111-117.

BRAYTON, K.A., COLLINS, N.E., VAN STRIJP, F. & ALLSOPP, B.A. 2003. Preparation of *Ehrlichia ruminantium* challenge material for quantifiable and reproducible challenge in mice and sheep. *Veterinary Pathology* **112**: 63-73.

BRAYTON, K.A., KAPPMEYER, L.S., HERNDON, D.R., DARK, M.J., TIBBALS, D.L., PALMER, G.H., MCGUIRE, T.C. & KNOWLES, D.P. Jr. 2005. Complete genome sequencing of *Anaplasma marginale* reveals that the surface is skewed to two superfamilies of outer membrane proteins. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 844-849.

BRENNAN, R.G. & MATTHEWS, B.W. 1989. The helix-turn-helix DNA binding motif. *Journal of Biological Chemistry* **264**: 1903-1906.

BROSCH, R., GORDON, S.V., BILLAULT, A., GARNIER, T., EIGLMEIER, K., SORAVITO, C., BARRELL, B.G. & COLE, S.T. 1998. Use of a *Mycobacterium tuberculosis* H37Rv bacterial

artificial chromosome library for genome mapping, sequencing, and comparative genomics. *Infection and Immunity* **66**: 2221-2229.

BROWN, W.C., MCELWAIN, T.F., RUEF, B.J., SUAREZ, C.E., SHKAP, V., CHITKO-MCKOWN, C.G., TUO, W., RICE-FICHT, A.C. & PALMER, G.H. 1996. *Babesia bovis* rhoptry-associated protein 1 is immunodominant for T helper cells of immune cattle and contains T-cell epitopes conserved among geographically distant *B. bovis* strains. *Infection and Immunity* **64**: 3341-3350.

BROWN, W.C., MCELWAIN, T.F., PALMER, G.H., CHANTLER, S.E. & ESTES, D.M. 1999. Bovine CD4(+) T-lymphocyte clones specific for rhoptry-associated protein 1 of *Babesia bigemina* stimulate enhanced immunoglobulin G1 (IgG1) and IgG2 synthesis. *Infection and Immunity* **67**: 155-164.

BROWN, J.S., OGUNNIYI, A.D., WOODROW, M.C., HOLDEN, D.W. & PATON, J.C. 2001. Immunization with components of two iron uptake ABC transporters protects mice against systemic *Streptococcus pneumoniae* infection. *Infection and Immunity* **69**: 6702-6706.

BYROM, B., BARBET, A.F., OBWOLO, M. & MAHAN, S.M. 2000. CD8+ T cell knockout mice are less susceptible to *Cowdria ruminantium* infection than athymic CD4+ T cell knockout, and normal C57BL/6 mice. *Veterinary Parasitology* **93**: 159-172.

CAMEJO, A., BUCHRIESER, C., COUVÉ, E., CARVALHO, F., REIS, O., FERREIRA, P., SOUSA, S., COSSART, P. & CABANES, D. 2009. *In vivo* transcriptional profiling of *Listeria monocytogenes* and mutagenesis identify new virulence factors involved in infection. *PLoS Pathogens* **5**: e1000449.

CAMUS, E. & BARRÉ, N. 1987. Diagnosis of heartwater in the live animal: experiences with goats in Guadeloupe. *Onderstepoort Journal of Veterinary Research* **54**: 291-294.

CAMUS, E. & BARRÉ, N. 1995. Vector situation of tick-borne diseases in the Caribbean Islands. *Veterinary Parasitology* **57**: 167-176.

CARVER, T.J., RUTHERFORD, K.M., BERRIMAN, M., RAJANDREAM, M.A., BARRELL, B.G. & PARKHILL, J. 2005. ACT: the Artemis Comparison Tool. *Bioinformatics* **21**: 3422-3423.

CATUREGLI, P., ASANOVICH, K.M., WALLS, J.J., BAKKEN, J.S., MADIGAN, J.E., POPOV, V.L. & DUMLER, J.S. 2000. *ankA*: an *Ehrlichia phagocytophila* group gene encoding a cytoplasmic protein antigen with ankyrin repeats. *Infection and Immunity* **68**: 5277-5283.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans:* a platform for investigating biology. *Science* **282**: 2012-2018.

CERRETTI, D.P., DEAN, D., DAVIS, G.R., BEDWELL, D.M. & NOMURA, M. 1983. The *spc* ribosomal protein operon of *Escherichia coli*: sequence and cotranscription of the ribosomal protein genes and a protein export gene. *Nucleic Acids Research* **11**: 2599-2616.

CHABALGOITY, J.A., BAZ, A., RIAL, A. & GRILLE, S. 2007. The relevance of cytokines for development of protective immunity and rational design of vaccines. *Cytokine & Growth Factor* **18**: 195-207.

CHAKRAVARTI, D.N., FISKE, M.J., FLETCHER, L.D. & ZAGURSKY, R.J. 2001. Application of genomics and proteomics for identification of bacterial gene products as potential vaccine candidates. *Vaccine* **19**: 601-612.

CHAPLIN, P.J., DE ROSE, R., BOYLE, J.S., MCWATERS, P., KELLY, J., TENNENT, J.M., LEW, A.M. & SCHEERLINCK, J.P. 1999. Targeting improves the efficacy of a DNA vaccine against *Corynebacterium pseudotuberculosis* in sheep. *Infection and Immunity* **67**: 6434-6438.

CHEN, J.M., COOPER, D.N., CHUZHANOVA, N., FÉREC, C. & PATRINOS, G.P. 2007. Gene conversion: mechanisms, evolution and human disease. *Nature Reviews Genetics* **8**: 762-775.

CHOO, K.H., TAN, T.W. & RANGANATHAN, S. 2009. A comprehensive assessment of N-terminal signal peptides prediction methods. *BMC Bioinformatics* **10**: S2.

CHRISTIE, P.J. 2001. Type IV secretion: intercellular transfer of macromolecules by systems ancestrally related to conjugation machines. *Molecular Microbiology* **40**: 294-305.

CITTI, C., KIM, M.F. & WISE, K.S. 1997. Elongated versions of Vlp surface lipoproteins protect *Mycoplasma hyorhinis* escape variants from growth-inhibiting host antibodies. *Infection and Immunity* **65**: 1773-1785.

COLLINS, N.E., PRETORIUS, A., VAN KLEEF, M., BRAYTON, K.A., ALLSOPP, M.T., ZWEYGARTH, E. & ALLSOPP, B.A. 2003. Development of improved attenuated and nucleic acid vaccines for heartwater. *Developments in Biologicals* **114**: 121-136.

COLLINS, N.E., LIEBENBERG, J., DE VILLIERS, E.P., BRAYTON, K.A., LOUW, E., PRETORIUS, A., FABER, F.E., VAN HEERDEN, H., JOSEMANS, A., VAN KLEEF, M., *et al.* 2005. The genome of the heartwater agent *Ehrlichia ruminantium* contains multiple tandem repeats of actively variable copy number. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 838-843.

CORE, L. & PEREGO, M. 2003. TPR-mediated interaction of RapC with ComA inhibits response regulator-DNA binding for competence development in *Bacillus subtilis*. *Molecular Microbiology* **49**: 1509-1522.

COSSART, P. & ARCHAMBAUD, C. 2009. The bacterial pathogen *Listeria monocytogenes*: an emerging model in prokaryotic transcriptomics. *Journal of Biology* **8**: 107.

COWDRY, E.V. 1925a. Studies on the etiology of heartwater. I. Observation of a *Rickettsia*, *Rickettsia ruminantium* in the tissues of infected animals. *Journal of Experimental Medicine* **42**: 231-252.

COWDRY, E.V. 1925b. Studies on the etiology of heartwater. II. *Rickettsia ruminantium* in the tissues of ticks transmitting the disease. *Journal of Experimental Medicine* **42**: 253-274.

COWDRY, E.V. 1926. Studies on the etiology of heartwater. III. The multiplication of *Rickettsia ruminantium* within the endothelial cells of infected animals and their discharge into the circulation. *Journal of Experimental Medicine* **44**: 803-814.

CROCQUET-VALDES, P.A., MCBRIDE, J.W., YU, X.J. & WALKER, D.H. 2003. Molecular characterization of the 28 kDa multigene locus of *Ehrlichia muris*. *Annals of the New York Academy of Sciences* **990**: 714-716.

CURASSON, M.G. 1943. Traité de protozoologie veterinaire et comparée. Trypanosomes, Volume I. Paris: Vigot Frères, pp 267-279.

DAS, A. & LJUNGDAHL, L.G. 2003. *Clostridium pasteurianum* F_1F_o ATP synthase: operon, composition, and some properties. *Journal of Bacteriology* **185**: 5527-5535.

DAS, A. & PAZOUR, G.J. 1989. Delineation of the regulatory region sequences of *Agrobacterium tumefaciens* virB operon. *Nucleic Acids Research* **17**: 4541-4550.

D'AURIA, G., JIMÉNEZ, N., PERIS-BONDIA, F., PELAZ, C., LATORRE, A. & MOYA, A. 2008. Virulence factor rtx in *Legionella pneumophila*, evidence suggesting it is a modular multifunctional protein. *BMC Genomics* **9**: 14.

DAVIS, H.L., MICHEL, M.L., MANCINI, M., SCHLEEF, M. & WHALEN, R.G. 1994. Direct gene transfer in skeletal muscle: plasmid DNA-based immunization against the hepatitis B virus surface antigen. *Vaccine* **12**: 1503-1509.

DEBOUCK, C. & GOODFELLOW, P.N. 1999. DNA microarrays in drug discovery and development. *Nature Genetics* **21**: 48-50.

DECKERS-HEBESTREIT, G. & ALTENDORF, K. 1996. The F_0F_1 -type ATP synthases of bacteria: structure and function of the F_0 complex. *Annual Review of Microbiology* **50**: 791-824.

DE LA FUENTE, J., GARCIA-GARCIA, J.C., BLOUIN, E.F. & KOCAN, K.M. 2003. Characterization of the functional domain of major surface protein 1a involved in adhesion of the rickettsia *Anaplasma marginale* to host cells. *Veterinary Microbiology* **91**: 265-283.

DE LA FUENTE, J., GARCIA-GARCIA, J.C., BARBET, A.F., BLOUIN, E.F. & KOCAN, K.M. 2004. Adhesion of outer membrane proteins containing tandem repeats of *Anaplasma* and *Ehrlichia* species (Rickettsiales: Anaplasmataceae) to tick cells. *Veterinary Microbiology* **98**: 313-322.

DE LA FUENTE, J., KOCAN, K.M., BLOUIN, E.F., ZIVKOVIC, Z., NARANJO, V., ALMAZÁN, C., ESTEVES, E., JONGEJAN, F., DAFFRE, S. & MANGOLD, A.J. 2010. Functional genomics and evolution of tick-*Anaplasma* interactions and vaccine development. *Veterinary Parasitology* **167**: 175-186.

DELCHER, A.L., HARMON, D., KASIF, S., WHITE, O. & SALZBERG, S.L. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Research* **27**: 4636-4641.

DEMEREC, M., ADELBERG, E.A., CLARK, A.J. & HARTMAN, P.E. 1966. A proposal for a uniform nomenclature in bacterial genetics. *Genetics* **54**: 61-76.

DE VILLIERS, E.P., BRAYTON, K.A., ZWEYGARTH, E. & ALLSOPP, B.A. 2000. Genome size and genetic map of *Cowdria ruminantium*. *Microbiology* **146**: 2627-2634.

DIGUISTINI, S., LIAO, N.Y., PLATT, D., ROBERTSON, G., SEIDEL, M., CHAN, S.K., DOCKING, T.R., BIROL, I., HOLT, R.A., HIRST, M., *et al*. 2009. *De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biology* **10**: R94.

DIXON, R.W. 1898. Heartwater experiments. *Agricultural Journal of the Cape of Good Hope* **12**: 754-760.

DOHM, J.C., LOTTAZ, C., BORODINA, T. & HIMMELBAUER, H. 2007. SHARCGS, a fast and highly accurate short-read assembly algorithm for *de novo* genomic sequencing. *Genome Research* **17**: 1697-1706.

DOOLITTLE, R.F. 1981. Similar amino acid sequences: chance or common ancestry? *Science* **214**: 149-159.

DORIA-ROSE, N.A. & HAIGWOOD, N.L. 2003. DNA vaccine strategies: candidates for immune modulation and immunization regimens. *Methods* **31**: 207-216.

DORO, F., LIBERATORI, S., RODRÍGUEZ-ORTEGA, M.J., RINAUDO, C.D., ROSINI, R., MORA, M., SCARSELLI, M., ALTINDIS, E., D'AURIZIO, R., STELLA, M., *et al*. 2009. Surfome analysis as a fast track to vaccine discovery: identification of a novel protective antigen for Group B *Streptococcus* hypervirulent strain COH1. *Molecular & Cellular Proteomics* **8**: 1728-1737.

DOYLE, C.K., ZHANG, X., POPOV, V.L. & MCBRIDE, J.W. 2005. An immunoreactive 38-kilodalton protein of *Ehrlichia canis* shares structural homology and iron-binding capacity with the ferric ion-binding protein family. *Infection and Immunity* **73**: 62-69.

DREW, D.R., BOYLE, J.S., LEW, A.M., LIGHTOWLERS, M.W., CHAPLIN, P.J. & STRUGNELL, R.A. 2001. The comparative efficacy of CTLA-4 and L-selectin targeted DNA vaccines in mice and sheep. *Vaccine* **19**: 4417-4428.

DROEGE, M. & HILL, B. 2008. The Genome Sequencer FLX System—longer reads, more applications, straight forward bioinformatics and more complete data sets. *Journal of Biotechnology* **136**: 3-10.

DUMLER, J.S., BARBET, A.F., BEKKER, C.P.J., DASCH, G.A., PALMER, G.H., RAY, S.C., RIKIHISA, Y. & RURANGIRWA, F.R. 2001. Reorganisation of genera in the families Rickettsiaceae and Anaplasmataceae in the order Rickettsiales: unification of some species of *Ehrlichia* and *Anaplasma*, *Cowdria* with *Ehrlichia* and *Ehrlichia* with *Neorickettsia*, descriptions of six new species combinations and designation of *Ehrlichia equi* and HE agent as subjective synonyms of *Ehrlichia phagocytophila*. *International Journal of Systematic and Evolutionary Microbiology* **51**: 2145-2165.

DUNHAM, S.P. 2002. The application of nucleic acid vaccines in veterinary medicine. *Research in Veterinary Science* **73**: 9-16.

DU PLESSIS, J.L. & KÜMM, N.A.L. 1971. The passage of *Cowdria ruminantium* in mice. *Journal of the South African Veterinary Medical Association* **42**: 217-221.

DU PLESSIS, J.L. 1985. A method for determining the *Cowdria ruminantium* infection rate of *Amblyomma hebraeum*: effects in mice injected with tick homogenates. *Onderstepoort Journal of Veterinary Research* **52**: 55-61.

DU PLESSIS, J.L. & MALAN, L. 1987. The non-specific resistance of cattle to heartwater. *Onderstepoort Journal of Veterinary Research* **54**: 333-336.

DU PLESSIS, J.L., VAN GAS, L., OLIVIER, J.A. & BEZUIDENHOUT, J.D. 1989. The heterogenicity of *Cowdria ruminantium* isolates: cross immunity and serology in sheep and pathogenicity to mice. *Onderstepoort Journal of Veterinary Research* **56**: 195-201.

DU PLESSIS, J.L., BERCHE, P. & VAN GAS, L. 1991. T cell-mediated immunity to *Cowdria ruminantium* in mice: the protective role of Lyt-2+ T cells. *Onderstepoort Journal of Veterinary Research* **58**: 171-179.

DU PLESSIS, J.L., BEZUIDENHOUT, J.D., BRETT, M.S., CAMUS, E., JONGEJAN, F., MAHAN, S.M. & MARTINEZ, D. 1993. The sero-diagnosis of heartwater: a comparison of five tests. *Revue d'Élevage et de Médecine Vétérinaire des Pays Tropicaux* **46**: 123-129.

EDINGTON, A. 1898. Heartwater. *Agricultural Journal of the Cape of Good Hope* **12**: 748-754.

EDWARDS, A., VOSS, H., RICE, P., CIVITELLO, A., STEGEMANN, J., SCHWAGER, C., ZIMMERMANN, J., ERFLE, H., CASKEY, C.T. & ANSORGE, W. 1990. Automated DNA sequencing of the human HPRT locus. *Genomics* **6**: 593-608.

ESTEVES, I., MARTINEZ, D. & TOTTÉ, P. 2004. Identification of *Ehrlichia ruminantium* (Gardel strain) IFN-γ inducing proteins after vaccination with a killed vaccine. *Veterinary Microbiology* **100**: 233-240.

EWING, B., HILLIER, L., WENDL, M.C. & GREEN, P. 1998. Base-calling automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8**: 175-185.

FELEK, S., HUANG, H. & RIKIHISA, Y. 2003. Sequence and expression analysis of *virB9* of the type IV secretion system of *Ehrlichia canis* strains in ticks, dogs, and cultured cells. *Infection and Immunity* **71**: 6063-6067.

FELSHEIM, R.F., HERRON, M.J., NELSON, C.M., BURKHARDT, N.Y., BARBET, A.F., KURTTI, T.J. & MUNDERLOH, U.G. 2006. Transformation of *Anaplasma phagocytophilum*. *BMC Biotechnology* **6**: 42.

FELSHEIM, R.F., CHÁVEZ, A.S., PALMER, G.H., CROSBY, L., BARBET, A.F., KURTTI, T.J. & MUNDERLOH, U.G. 2010. Transformation of *Anaplasma marginale*. *Veterinary Parasitology* **167**: 167-174.

FENN, K. & BLAXTER, M. 2006. *Wolbachia* genomes: revealing the biology of parasitism and mutualism. *Trends in Parasitology* **22**: 60-65.

FINLAY, B.B. & FALKOW, S. 1997. Common themes in microbial pathogenicity revisited. *Microbiology and Molecular Biology Reviews* **61**: 136-169.

FLACH, E.J., WOODFORD, J.D., MORZARIA, S.P., DOLAN, T.T. & SHAMBWANA, I. 1990. Identification of *Babesia bovis* and *Cowdria ruminantium* on the island of Unguja, Zanzibar. *The Veterinary Record* **126**: 57-59.

FLEISCHMANN, R.D., ADAMS, M.D., WHITE, O., CLAYTON, R.A., KIRKNESS, E.F., KERLAVAGE, A.R., BULT, C.J., TOMB, J.F., DOUGHERTY, B.A., MERRICK, J.M., *et al.* 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**: 496-512.

FOSTER, J., GANATRA, M., KAMAL, I., WARE, J., MAKAROVA, K., IVANOVA, N., BHATTACHARYYA, A., KAPATRAL, V., KUMAR, S., POSFAI, J., *et al.* 2005. The *Wolbachia* genome of *Brugia malayi*: endosymbiont evolution within a human pathogenic nematode. *PLoS Biology* **3**: e121.

FRANGEUL, L., NELSON, K.E., BUCHRIESER, C., DANCHIN, A., GLASER, P. & KUNST, F. 1999. Cloning and assembly strategies in microbial genome projects. *Microbiology* **145**: 2625-2634.

FREDRICKS, D.N. 2006. Introduction to the Rickettsiales and other intracellular prokaryotes. In: The prokaryotes: A handbook on the biology of bacteria, 3rd ed. Edited by M. Dworkin, S. Falkow, E. Rosenberg, K.H. Schleifer and E. Stackebrandt. New York: Springer, pp 457-466.

FRISHMAN, D., MIRONOV, A., MEWES, H.W. & GELFAND, M. 1998. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. *Nucleic Acids Research* **26**: 2941-2947.

FRUTOS, R., VIARI, A., FERRAZ, C., MORGAT, A., EYCHENIÉ, S., KANDASSAMY, Y., CHANTAL, I., BENSAID, A., COISSAC, E., VACHIERY, N., *et al.* 2006. Comparative genomic analysis of three strains of *Ehrlichia ruminantium* reveals an active process of genome size plasticity. *Journal of Bacteriology* **188**: 2533-2542.

FRUTOS, R., VIARI, A., VACHIERY, N., BOYER, F. & MARTINEZ, D. 2007. *Ehrlichia ruminantium*: genomic and evolutionary features. *Trends in Parasitology* **23**: 414-419.

FU, X., FU, N., GUO, S., YAN, Z., XU, Y., HU, H., MENZEL, C., CHEN, W., LI, Y., ZENG, R. & KHAITOVICH, P. 2009. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* **10**: 161.

GALINDO, R.C., AYOUBI, P., GARCÍA-PÉREZ, A.L., NARANJO, V., KOCAN, K.M., GORTAZAR, C. & DE LA FUENTE, J. 2008. Differential expression of inflammatory and immune response genes in sheep infected with *Anaplasma phagocytophilum*. *Veterinary Immunology and Immunopathology* **126**: 27-34.

GARCIA-GARCIA, J.C., DE LA FUENTE, J., BELL-EUNICE, G., BLOUIN, E.F. & KOCAN, K.M. 2004. Glycosylation of *Anaplasma marginale* major surface protein 1a and its putative role in adhesion to tick cells. *Infection and Immunity* **72**: 3022-3030.

GARDNER, M.J., HALL, N., FUNG, E., WHITE, O., BERRIMAN, M., HYMAN, R.W., CARLTON, J.M., PAIN, A., NELSON, K.E., BOWMAN, S., *et al.* 2002. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**: 498-511.

GARDY, J.L., LAIRD, M.R., CHEN, F., REY, S., WALSH, C.J., ESTER, M. & BRINKMAN, F.S.L. 2005. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localisation and insights gained from comparative proteome analysis. *Bioinformatics* **21**: 617-623.

GIOVANNONI, S.J., TRIPP, H.J., GIVAN, S., PODAR, M., VERGIN, K.L., BAPTISTA, D., BIBBS, L., EADS, J., RICHARDSON, T.H., NOORDEWIER, M., *et al.* 2005. Genome streamlining in a cosmopolitan oceanic bacterium. *Science* **309**: 1242-1245.

GIULIANI, M.M., ADU-BOBIE, J., COMANDUCCI, M., ARICÒ, B., SAVINO, S., SANTINI, L., BRUNELLI, B., BAMBINI, S., BIOLCHI, A., CAPECCHI, B., *et al.* 2006. A universal vaccine for serogroup B meningococcus. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 10834-10839.

GOLDBERG, S.M., JOHNSON, J., BUSAM, D., FELDBLYUM, T., FERRIERA, S., FRIEDMAN, R., HALPERN, A., KHOURI, H., KRAVITZ, S.A., LAURO, F.M., *et al.* 2006. A Sanger/ pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 11240-11245.

GRANDI, G. 2001. Antibacterial vaccine design using genomics and proteomics. *Trends in Biotechnology* **19**: 181-188.

GRANDI, G. 2003. Rational antibacterial vaccine design through genomic technologies. *International Journal for Parasitology* **33**: 615-620.

GRIFANTINI, R., BARTOLINI, E., MUZZI, A., DRAGHI, M., FRIGIMELICA, E., BERGER, J., RATTI, G., PETRACCA, R., GALLI, G., AGNUSDEI, M., *et al.* 2002. Previously unrecognized vaccine

candidates against group B meningococcus identified by DNA microarrays. *Nature Biotechnology* **20**: 914-921.

GUPTA, R.S. & MOK, A. 2007. Phylogenomics and signature proteins for the alpha Proteobacteria and its main groups. *BMC Microbiology* **7**: 106.

HAIG, D.A., ALEXANDER, R.A. & WEISS, K.E. 1954. Treatment of heartwater with tetramycin. *Journal of the South African Veterinary Medical Association* **25**: 45-48.

HALL, N. 2007. Advanced sequencing technologies and their wider impact in microbiology. *The Journal of Experimental Biology* **209**: 1518-1525.

HARDING, C.V., RAMACHANDRA, L. & WICK, M.J. 2003. Interaction of bacteria with antigen presenting cells: influences on antigen presentation and antibacterial immunity. *Current Opinion in Immunology* **15**: 112-119.

HARRIS, T.D., BUZBY, P.R., BABCOCK, H., BEER, E., BOWERS, J., BRASLAVSKY, I., CAUSEY, M., COLONELL, J., DIMEO, J., EFCAVITCH, J.W., *et al*. 2008. Single-molecule DNA sequencing of a viral genome. *Science* **320**: 106-109.

HARRISON, R.G. 2000. Expression of soluble heterologous proteins via fusion with NusA protein. *inNovations* **11**: 4-7.

HATCH, T.P. 1999. Developmental biology. In: *Chlamydia:* intracellular biology, pathogenesis, and immunity. Edited by R.S. Stephens. Washington, D.C.: American Society for Microbiology, pp 29-67.

HICKLING, J.K. 1998. Measuring human T-lymphocyte function. *Expert Reviews in Molecular Medicine* **1998**: 1-20.

HIGGINS, C.F. 2001. ABC transporters: physiology, structure and mechanism – an overview. *Research in Microbiology* **152**: 205-210.

HILLIER, L.W., MARTH, G.T., QUINLAN, A.R., DOOLING, D., FEWELL, G., BARNETT, D., FOX, P., GLASSCOCK, J.I., HICKENBOTHAM, M., HUANG, W., *et al*. 2008. Whole-genome sequencing and variant discovery in *C. elegans*. *Nature Methods* **5**: 183-188.

HOLT, K.E., PARKHILL, J., MAZZONI, C.J., ROUMAGNAC, P., WEILL, F.X., GOODHEAD, I., RANCE, R., BAKER, S., MASKELL, D.J., WAIN, J., *et al*. 2008. High-throughput sequencing

provides insights into genome variation and evolution in *Salmonella* Typhi. *Nature Genetics* **40**: 987-993.

HOOD, D.W., DEADMAN, M.E., JENNINGS, M.P., BISERCIC, M., FLEISCHMANN, R.D., VENTER, J.C. & MOXON, E.R. 1996. DNA repeats identify novel virulence genes in *Haemophilus influenzae*. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 11121-11125.

HOTOPP, J.C., LIN, M., MADUPU, R., CRABTREE, J., ANGIUOLI, S.V., EISEN, J., SESHADRI, R., REN, Q., WU, M., UTTERBACK, T.R. *et al.* 2006. Comparative genomics of emerging human ehrlichiosis agents. *PLoS Genetics* **2**: e21.

HUGHES, D. 2000a. Evaluating genome dynamics: the constraints on rearrangements within bacterial genomes. *Genome Biology* **1**: review 0006.1-0006.8.

HUGHES, D. 2000b. Co-evolution of the *tuf* genes links gene conversion with the generation of chromosomal inversions. *Journal of Molecular Biology* **297**: 355-364.

HUTCHEON, D. 1900. History of heartwater. *Agricultural Journal of the Cape of Good Hope* **17**: 410-417.

HUYGEN, K. 2003. On the use of DNA vaccines for the prophylaxis of mycobacterial diseases. *Infection and Immunity* **71**: 1613-1621.

IBBA, M., CURNOW, A.W. & SOLL, D. 1997. Aminoacyl-tRNA synthesis: divergent routes to a common goal. *Trends in Biochemical Sciences* **22**: 39-42.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860-921.

IOANNIDIS, P., HOTOPP, J.C., SAPOUNTZIS, P., SIOZIOS, S., TSIAMIS, G., BORDENSTEIN, S.R., BALDO, L., WERREN, J.H. & BOURTZIS, K. 2007. New criteria for selecting the origin of DNA replication in *Wolbachia* and closely related bacteria. *BMC Genomics* **8**: 182.

JIANG, G., CHAROENVIT, Y., MORENO, A., BARACEROS, M.F., BANANIA, G., RICHIE, N., ABOT, S., GANESHAN, H., FALLARME, V., PATTERSON, N.B., *et al.* 2007. Induction of multi-antigen multi-stage immune responses against *Plasmodium falciparum* in rhesus monkeys, in the absence of antigen interference, with heterologous DNA prime/poxvirus boost immunization. *Malaria Journal* **6**: 135.

JONGEJAN, F. 1991. Protective immunity to heartwater (*Cowdria ruminantium* infection) is acquired after vaccination with in vitro-attenuated rickettsiae. *Infection and Immunity* **59**: 729-731.

JOSEMANS, A.I. & ZWEYGARTH, E. 2002. Amino acid content of cell cultures infected with *Cowdria ruminantium* propagated in a protein-free medium. *Annals of the New York Academy of Sciences* **969**: 141-146.

JUHAS, M., CROOK, D.W. & HOOD, D.W. 2008. Type IV secretion systems: tools of bacterial horizontal gene transfer and virulence. *Cellular Microbiology* **10**: 2377-2386.

JURINKE, C., VAN DEN BOOM, D., CANTOR, C.R. & KOSTER, H. 2002. The use of MassARRAY technology for high throughput genotyping. *Advances in Biochemical Engineering / Biotechnology* **77**: 57-74.

KÄLL, L., KROGH, A. & SONNHAMMER, E.L.L. 2004. A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology* **338**: 1027-1036.

KALMAN, S., MITCHELL, W., MARATHE, R., LAMMEL, C., FAN, J., HYMAN, R.W., OLINGER, L., GRIMWOOD, J., DAVIS, R.W. & STEPHENS, R.S. 1999. Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nature Genetics* **21**: 385-389.

KANEHISA, M. & GOTO, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28**: 27-30.

KANEKO, T., NAKAMURA, Y., SATO, S., ASAMIZU, E., KATO, T., SASAMOTO, S., WATANABE, A., IDESAWA, K., ISHIKAWA, A., KAWASHIMA, K., *et al*. 2000. Complete genome structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. *DNA Research* **7**: 331-338.

KANYARI, P.W. & KAGIRA, J. 2000. The role of parasitic diseases as causes of mortality in cattle in a high potential area of central Kenya: a quantitative analysis. *Onderstepoort Journal of Veterinary Research* **67**: 157-161.

KASIANOWICZ, J.J., BRANDIN, E., BRANTON, D. & DEAMER, D.W. 1996. Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences of the United States of America* **93**: 13770-13773.

KATHAPERUMAL, K., KUMANAN, V., MCDONOUGH, S., CHEN, L.H., PARK, S.U., MOREIRA, M.A., AKEY, B., HUNTLEY, J., CHANG, C.F. & CHANG, Y.F. 2009. Evaluation of immune responses and protective efficacy in a goat model following immunization with a coctail of

recombinant antigens and a polyprotein of *Mycobacterium avium* subsp. *paratuberculosis*. *Vaccine* **27**: 123-135.

KAUSHIK, D.K. & SEHGAL, D. 2008. Developing antibacterial vaccines in genomics and proteomics era. *Scandinavian Journal of Immunology* **67**: 544-552.

KENNEDY, N.J., SPITHILL, T.W., TENNENT, J., WOOD, P.R. & PIEDRAFITA, D. 2006. DNA vaccines in sheep: CTLA-4 mediated targeting and CpG motifs enhance immunogenicity in a DNA prime/protein boost strategy. *Vaccine* **24**: 970-979.

KOCAN, K.M. & BEZUIDENHOUT, J.D. 1987. Morphology and development of *Cowdria ruminantium* in *Amblyomma* ticks. *Onderstepoort Journal of Veterinary Research* **54**: 177-182.

KOCAN, K.M., BEZUIDENHOUT, J.D. & HART, A. 1987. Ultrastructural features of *Cowdria ruminantium* in midgut epithelial cells and salivary glands of nymphal *Amblyomma hebraeum*. *Onderstepoort Journal of Veterinary Research* **54**: 87-92.

KOCK, N.D., VAN VLIET, A.H.M., CHARLTON, K. & JONGEJAN, F. 1995. Detection of *Cowdria ruminantium* in blood and bone marrow samples from clinically normal, free-ranging Zimbabwean wild ungulates. *Journal of Clinical Microbiology* **33**: 2501-2504.

KODAMA, K., KAWAMURA, S., YASUKAWA, M. & KOBAYASHI, Y. 1987. Establishment and characterization of a T-cell line specific for *Rickettsia tsutsugamushi*. *Infection and Immunity* **55**: 2490-2495.

KOIDE, T., ZAINI, P.A., MOREIRA, L.M., VÊNCIO, R.Z., MATSUKUMA, A.Y., DURHAM, A.M., TEIXEIRA, D.C., EL-DORRY, H., MONTEIRO, P.B., DA SILVA, A.C., *et al*. 2004. DNA microarray-based genome comparison of a pathogenic and a nonpathogenic strain of *Xylella fastidiosa* delineates genes important for bacterial virulence. *Journal of Bacteriology* **186**: 5442-5449.

KOLPAKOV, R., BANA, G. & KUCHEROV, G. 2003. mreps: Efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Research* **31**: 3672-3678.

KROGH, A., LARSSON, B., VON HEIJNE, G. & SONNHAMMER, E.L.L. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology* **305**: 567-580.

KUMAGAI, Y., HUANG, H. & RIKIHISA, Y. 2008. Expression and porin activity of P28 and OMP-1F during intracellular *Ehrlichia chaffeensis* development. *Journal of* Bacteriology **190**: 3597-3605.

KUNST, F., OGASAWARA, N., MOSZER, I., ALBERTINI, A.M., ALLONI, G., AZEVEDO, V., BERTERO, M.G., BESSIÈRES, P., BOLOTIN, A., BORCHERT, S., *et al.* 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature* **390**: 249-256.

LANGDON, S.P. 2004. Cell culture contamination: an overview. *Methods in Molecular Medicine* **88**: 309-317.

LAWSON, M.J., JIAO, J., FAN, W. & ZHANG, L. 2009. A pattern analysis of gene conversion literature. *Comparative and Functional Genomics* **2009**: 761512.

LEUNG, W.H., MENG, Z.Q., HUI, G. & HO, W.K. 2004. Expression of an immunologically reactive merozoite surface protein ($MSP\text{-}1_{42}$) in *E. coli*. *Biochimica et Biophysica Acta* **1675**: 62-70.

LEVINSON, G. & GUTMAN, G.A. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molecular Biology and Evolution* **4**: 203-221.

LIN, J., HUANG, S. & ZHANG, Q. 2002. Outer membrane proteins: key players for bacterial adaptation in host niches. *Microbes and Infection* **4**: 325-331.

LIN, M. & RIKIHISHA, Y. 2003. *Ehrlichia chaffeensis* and *Anaplasma phagocytophilum* lack genes for lipid A biosynthesis and incorporate cholesterol for their survival. *Infection and Immunity* **71**: 5324-5331.

LIN, M., DEN DULK-RAS, A., HOOYKAAS, P.J. & RIKIHISA, Y. 2007. *Anaplasma phagocytophilum* AnkA secreted by type IV secretion system is tyrosine phosphorylated by Abl-1 to facilitate infection. *Cellular Microbiology* **9**: 2644-2657.

LIN, M., ZHANG, C., GIBSON, K. & RIKIHISA, Y. 2009. Analysis of complete genome sequence of *Neorickettsia risticii*: causative agent of Potomac horse fever. *Nucleic Acids Research* **37**: 6076-6091.

LOBRY, J.R. 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution* **13**: 660-665.

LOGAN, L.L., WHYARD, T.C., QUINTERO, J.C. & MEBUS, C.A. 1987. The development of *Cowdria ruminantium* in neutrophils. *Onderstepoort Journal of Veterinary Research* **54**: 197-204.

LONG, S.W., WHITWORTH, T.J., WALKER, D.H. & YU, X.J. 2005. Overcoming barriers to the transformation of the genus *Ehrlichia*. *Annals of the New York Academy of Sciences* **1063**: 403-410.

LOPEZ, J.E., SIEMS, W.F., PALMER, G.H., BRAYTON, K.A., MCGUIRE, T.C., NORIMINE, J. & BROWN, W.C. 2005. Identification of novel antigenic proteins in a complex *Anaplasma marginale* outer membrane immunogen by mass spectrometry and genomic mapping. *Infection and Immunity* **73**: 8109-8118.

LOPEZ, J.E., PALMER, G.H., BRAYTON, K.A., DARK, M.J., LEACH, S.E. & BROWN, W.C. 2007. Immunogenicity of *Anaplasma marginale* type IV secretion system proteins in a protective outer membrane vaccine. *Infection and Immunity* **75**: 2333-2342.

LORENZEN, N. & LAPATRA, S.E. 2005. DNA vaccines for aquacultured fish. *Revue Scientifique et Technique* **24**: 201-213.

LOUNSBURY, C.P. 1900. Tick-heartwater experiment. *Agricultural Journal of the Cape of Good Hope* **16**: 682-687.

LOUW, E., BRAYTON, K.A., COLLINS, N.E., PRETORIUS, A., VAN STRIJP, F. & ALLSOPP, B.A. 2002. Sequencing of a 15-kb *Ehrlichia ruminantium* clone and evaluation of the *cpg1* open reading frame for protection against heartwater. *Annals of the New York Academy of Sciences* **969**: 147-150.

LOWE, T.M. & EDDY, S.R. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* **25**: 955-964.

LUDWIG, W. & KLENK, H.P. 2001. Overview: A phylogenetic backbone and taxonomic framework for prokaryotic systamatics. In: Bergey's manual of systematic bacteriology, 2nd ed. Edited by D.R. Boone and R.W. Castenholz. Berlin: Springer-Verlag, pp 49-65.

LUO, D., NI, B., LI, P., SHI, W., ZHANG, S., HAN, Y., MAO, L., HE, Y., WU, Y. & WANG, X. 2006. Protective immunity elicited by a divalent DNA vaccine encoding both the L7/L12 and Omp16 genes of *Brucella abortus* in BALB/c mice. *Infection and Immunity* **74**: 2734-2741.

LUO, T., ZHANG, X., WAKEEL, A., POPOV, V.L. & MCBRIDE, J.W. 2008. A variable-length PCR target protein of *Ehrlichia chaffeensis* contains major species-specific antibody epitopes in acidic serine-rich tandem repeats. *Infection and Immunity* **76**: 1572-1580.

LUO, T., ZHANG, X. & MCBRIDE, J.W. 2009. Major species-specific antibody epitopes of the *Ehrlichia chaffeensis* p120 and *E. canis* p140 orthologs in surface-exposed tandem repeat regions. *Clinical and Vaccine Immunology* **16**: 982-990.

MADABHUSHI, R.S. 1998. Separation of 4-color DNA sequencing extension products in noncovalently coated capillaries using low viscosity polymer solutions. *Electrophoresis* **19**: 224-230.

MAHAN, S.M., ANDREW, H.R., TEBELE, N., BURRIDGE, M.J. & BARBET, A.F. 1995. Immunization of sheep against heartwater with inactivated *Cowdria ruminantium*. *Research in Veterinary Science* **58**: 46-49.

MAHAN, S.M., PETER, T.F., SIMBI, B.H., KOCAN, K., CAMUS, E., BARBET, A.F. & BURRIDGE, M.J. 2000. Comparison of efficacy of American and African *Amblyomma* ticks as vectors of heartwater (*Cowdria ruminantium*) infection by molecular analyses and transmission trials. *Journal of Parasitology* **86**: 44-49.

MAHAN, S.M., SMITH, G.E., KUMBULA, D., BURRIDGE, M.J. & BARBET, A.F. 2001. Reduction in mortality from heartwater in cattle, sheep and goats exposed to field challenge using an inactivated vaccine. *Veterinary Parasitology* **97**: 295-308.

MAILLARD, J.C. & MAILLARD, N. 1998. Historique du peuplement bovin et de l'introduction de la tique *Amblyomma variegatum* dans les îles françaises des Antilles: Synthèse bibliographique. *Ethnozootechnie* **1**: 19-36.

MAIONE, D., MARGARIT, I., RINAUDO, C.D., MASIGNANI, V., MORA, M., SCARSELLI, M., TETTELIN, H., BRETTONI, C., IACOBINI, E.T., ROSINI, R., *et al*. 2005. Identification of a universal Group B *Streptococcus* vaccine by multiple genome screen. *Science* **309**: 148-150.

MANNING, S.D., MOTIWALA, A.S., SPRINGMAN, A.C., QI, W., LACHER, D.W., OUELLETTE, L.M., MLADONICKY, J.M., SOMSEL, P., RUDRIK, J.T., DIETRICH, S.E., *et al*. 2008. Variation in virulence among clades of *Escherichia coli* O157:H7 associated with disease outbreaks. *Proceedings of the National Academy of Sciences of the United States of America* **105**: 4868-4873.

MARGUERAT, S. & BÄHLER, J. 2009. RNA-seq: from technology to biology. *Cellular and Molecular Life Sciences* **67**: 569-79.

MARGULIES, M., EGHOLM, M., ALTMAN, W.E., ATTIYA, S., BADER, J.S., BEMBEN, L.A., BERKA, J., BRAVERMAN, M.S., CHEN, Y.J., CHEN, Z., *et al*. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**: 376-380.

MARIOTTI, E., MIRABELLI, P., DI NOTO, R., FORTUNATO, G. & SALVATORE, F. 2008. Rapid detection of mycoplasma in continuous cell lines using a selective biochemical test. *Leukemia Research* **32**: 323-326.

MARTINEZ, D., MAILLARD, J.C., COISNE, S., SHEIKBOUDOU, C. & BENSAID, A. 1994. Protection of goats against heartwater acquired by immunisation with inactivated elementary bodies of *Cowdria ruminantium*. *Veterinary Immunology and Immunopathology* **41**: 153-163.

MAVROMATIS, K., DOYLE, C.K., LYKIDIS, A., IVANOVA, N., FRANCINO, M.P., CHAIN, P., SHIN, M., MALFATTI, S., LARIMER, F., COPELAND, A., *et al.* 2006. The genome of the obligately intracellular bacterium *Ehrlichia canis* reveals themes of complex membrane structure and immune evasion strategies. *Journal of Bacteriology* **188**: 4015-4023.

MAXAM, A.M. & GILBERT, W. 1977. A new method for sequencing DNA. *Proceedings of the National Academy of Sciences of the United States of America* **74**: 560-564.

MCBRIDE, J.W., YU, X.J. & WALKER, D.H. 2000. Glycosylation of homologous immunodominant proteins of *Ehrlichia chaffeensis* and *Ehrlichia canis*. *Infection and Immunity* **68**: 13-18.

MEEUS, P.F. & BARBET, A.F. 2001. Ingenious gene generation. *Trends in Microbiology* **9**: 353-355.

MEEUS, P.F., BRAYTON, K.A., PALMER, G.H. & BARBET, A.F. 2003. Conservation of a gene conversion mechanism in two distantly related paralogues of *Anaplasma marginale*. *Molecular Microbiology* **47**: 633-643.

MERRELL, D.S., BUTLER, S.M., QADRI, F., DOLGANOV, N.A., ALAM, A., COHEN, M.B., CALDERWOOD, S.B., SCHOOLNIK, G.K. & CAMILLI, A. 2002. Host induced epidemic spread of the cholera bacterium. *Nature* **417**: 642-645.

MIN, C.K., YANG, J.S., KIM, S., CHOI, M.S., KIM, I.S. & CHO, N.H. 2008. Genome-based construction of the metabolic pathways of *Orientia tsutsugamushi* and comparative analysis within the Rickettsiales order. *Comparative and Functional Genomics* **2008**: 623145.

MIYOSHI, S. & SHINODA, S. 2000. Microbial metalloproteases and pathogenesis. *Microbes and Infection* **2**: 91-98.

MONTIGIANI, S., FALUGI, F., SCARSELLI, M., FINCO, O., PETRACCA, R., GALLI, G., MARIANI, M., MANETTI, R., AGNUSDEI, M., CEVENINI, R., *et al*. 2002. Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae*. *Infection and Immunity* **70**: 368-379.

MORTAZAVI, A., WILLIAMS, B.A., MCCUE, K., SCHAEFFER, L. & WOLD, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**: 621-628.

MOSHKOVSKI, S.D. 1947. Comments by readers. *Science* **106**: 62.

MOSMANN, T.R. & SAD, S. 1996. The expanding universe of T-cell subsets: Th1, Th2 and more. *Immunology Today* **17**: 138-146.

MOXON, E.R., HOOD, D.W., SAUNDERS, N.J., SCHWEDA, E.K. & RICHARDS, J.C. 2002. Functional genomics of pathogenic bacteria. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **357**: 109-116.

MULLER KOBOLD, A., MARTINEZ, D., CAMUS, E. & JONGEJAN, F. 1992. Distribution of heartwater in the Caribbean determined on the basis of detection of antibodies to the conserved 32-kilodalton protein of *Cowdria ruminantium*. *Journal of Clinical Microbiology* **30**: 1870-1873.

MUZZI, A., MASIGNANI, V. & RAPPUOLI, R. 2007. The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials. *Drug Discovery Today* **12**: 429-439.

MWANGI, D.M., MAHAN, S.M., NYANJUI, J.K., TARACHA, E.L.N. & MCKEEVER, D.J. 1998. Immunization of cattle by infection with *Cowdria ruminantium* elicits T lymphocytes that recognise autologous infected endothelial cells and monocytes. *Infection and Immunity* **66**: 1855-1860.

MWANGI, D.M., MCKEEVER, D.J., NYANJUI, J.K., BARBET, A.F. & MAHAN, S.M. 2002. Immunisation of cattle against heartwater by infection with *Cowdria ruminantium* elicits T lymphocytes that recognise major antigenic proteins 1 and 2 of the agent. *Veterinary Immunology and Immunopathology* **85**: 23-32.

MYERS, G.S., PARKER, D., AL-HASANI, K., KENNAN, R.M., SEEMANN, T., REN, Q., BADGER, J.H., SELENGUT, J.D., DEBOY, R.T., TETTELIN, H., *et al.* 2007. Genome sequence and identification of candidate vaccine antigens from the animal pathogen *Dichelobacter nodosus*. *Nature Biotechnology* **25**: 569-575.

NAGALAKSHMI, U., WANG, Z., WAERN, K., SHOU, C., RAHA, D., GERSTEIN, M. & SNYDER, M. 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* **320**: 1344-1349.

NAKAMURA, Y., LEPPERT, M., O'CONNELL, P., WOLFF, R., HOLM, T., CULVER, M., MARTIN, C., FUJIMOTO, E., HOFF, M., KUMLIN, E., *et al.* 1987. Variable number of tandem repeat (VNTR) markers for human gene mapping. *Science* **235**: 1616-1622.

NAVARRE, W.W. & SCHNEEWIND, O. 1999. Surface proteins of gram-positive bacteria and mechanisms of their targeting to the cell wall envelope. *Microbiology and Molecular Biology Reviews* **63**: 174-229.

NEITZ, W.O. 1940. Uleron in the treatment of heartwater. *Journal of the South African Veterinary Medical Association* **11**: 15.

NEITZ, W.O. & ALEXANDER, R.A. 1941. The immunization of calves against heartwater. *Journal of the South African Veterinary Medical Association* **12**: 103-111.

NEITZ, W.O. & ALEXANDER, R.A. 1945. Immunization of cattle against heartwater and the control of the tick-borne diseases, redwater, gallsickness and heartwater. *Onderstepoort Journal of Veterinary Science and Animal Industry* **20**: 137-158.

NEITZ, W.O. 1964. Tick-borne diseases as a hazard in the rearing of calves in South Africa. *Bulletin - Office international des épizooties* **62**: 607-625.

NEITZ, W.O. 1968. Heartwater. *Bulletin - Office international des épizooties* **70**: 329-336.

NICHOLS, W.W., LEDWITH, B.J., MANAM, S.V. & TROILO, P.J. 1995. Potential DNA vaccine integration into host cell genome. *Annals of the New York Academy of Sciences* **772**: 30-39.

NIELSEN, H., ENGELBRECHT, J., BRUNAK, S. & VON HEIJNE, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Engineering* **10**: 1-6.

NIEMANN, H.H., SCHUBERT, W.D. & HEINZ, D.W. 2004. Adhesins and invasins of pathogenic bacteria: a structural view. *Microbes and Infection* **6**: 101-112.

NOVAES, E., DROST, D.R., FARMERIE, W.G., PAPPAS, G.J. Jr., GRATTAPAGLIA, D., SEDEROFF, R.R. & KIRST, M. 2008. High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. *BMC Genomics* **9**: 312.

NYIKA, A., MAHAN, S.M., BURRIDGE, M.J., MCGUIRE, T.C., RURANGIRWA, F. & BARBET, A.F. 1998. A DNA vaccine protects mice against the rickettsial agent *Cowdria ruminantium*. *Parasite Immunology* **20**: 111-119.

NYIKA, A., BARBET, A.F., BURRIDGE, M.J. & MAHAN, S.M. 2002. DNA vaccination with *map1* gene followed by protein boost augments protection against challenge with *Cowdria ruminantium*, the agent of heartwater. *Vaccine* **20**: 1215-1225.

OBEREM, P.T. & BEZUIDENHOUT, J.D. 1987. The production of heartwater vaccine. *Onderstepoort Journal of Veterinary Research* **54**: 485-488.

OGASAWARA, N. & YOSHIKAWA, H. 1992. Genes and their organization in the replication origin region of the bacterial chromosome. *Molecular Microbiology* **6**: 629-634.

OGATA, H., GOTO, S., SATO, K., FUJIBUCHI, W., BONO, H. & KANEHISA, M. 1999. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **27**: 29-34.

OGATA, H., AUDIC, S., BARBE, V., ARTIGUENAVE, F., FOURNIER, P.E., RAOULT, D. & CLAVERIE, J.M. 2000. Selfish DNA in protein-coding genes of Rickettsia. *Science* **290**: 347-350.

OGATA, H., RENESTO, P., AUDIC, S., ROBERT, C., BLANC, G., FOURNIER, P.E., PARINELLO, H., CLAVERIE, J.M. & RAOULT, D. 2005. The genome sequence of *Rickettsia felis* identifies the first putative conjugative plasmid in an obligate intracellular parasite. *PLoS Biology* **3**: e248.

OGATA, H., LA SCOLA, B., AUDIC, S., RENESTO, P., BLANC, G., ROBERT, C., FOURNIER, P.E., CLAVERIE, J.M. & RAOULT, D. 2006. Genome sequence of *Rickettsia bellii* illuminates the role of amoebae in gene exchanges between intracellular pathogens. *PLoS Genetics* **2**: e76.

OHASHI, N., UNVER, A., ZHI, N. & RIKIHISA, Y. 1998a. Cloning and characterization of multigenes encoding the immunodominant 30-kilodalton major outer membrane proteins of *Ehrlichia canis* and application of the recombinant protein for serodiagnosis. *Journal of Clinical Microbiology* **36**: 2671-2680.

OHASHI, N., ZHI, N., ZHANG, Y. & RIKIHISA, Y. 1998b. Immunodominant major outer membrane proteins of *Ehrlichia chaffeensis* are encoded by a polymorphic multigene family. *Infection and Immunity* **66**: 132-139.

OJCIUS, D.M., GACHELIN, G. & DANTRY-VARSAT, A. 1996. Presentation of antigens derived from microorganisms residing in host-cell vacuoles. *Trends in Microbiology* **4**: 53-59.

OÑATE, A.A., DONOSO, G., MORAGA-CID, G., FOLCH, H., CÉSPEDES, S. & ANDREWS, E. 2005. An RNA vaccine based on recombinant Semliki Forest virus particles expressing the Cu,Zn superoxide dismutase protein of *Brucella abortus* induces protective immunity in BALB/c mice. *Infection and Immunity* **73**: 3294-3300.

OSADA, N. & INNAN, H. 2008. Duplication and gene conversion in the *Drosophila melanogaster* genome. *PLoS Genetics* **4**: e1000305.

OZSOLAK, F., PLATT, A.R., JONES, D.R., REIFENBERGER, J.G., SASS, L.E., MCINERNEY, P., THOMPSON, J.F., BOWERS, J., JAROSZ, M. & MILOS, P.M. 2009. Direct RNA sequencing. *Nature* **461**: 814-818.

PAN, W., RAVOT, E., TOLLE, R., FRANK, R., MOSBACH, R., TÜRBACHOVA, I. & BUJARD, H. 1999. Vaccine candidate MSP-1 from *Plasmodium falciparum*: a redesigned 4917 bp polynucleotide

enables synthesis and isolation of full-length protein from *Escherichia coli* and mammalian cells. *Nucleic Acids Research* **27**: 1094-1103.

PARK, S.U., KATHAPERUMAL, K., MCDONOUGH, S., AKEY, B., HUNTLEY, J., BANNANTINE, J.P. & CHANG, Y.F. 2008. Immunization with a DNA vaccine cocktail induces a Th1 response and protects mice against *Mycobacterium avium* subsp. *paratuberculosis* challenge. *Vaccine* **26**: 4329-4337.

PARKHILL, J., WREN, B.W., MUNGALL, K., KETLEY, J.M., CHURCHER, C., BASHAM, D., CHILLINGWORTH, T., DAVIES, R.M., FELTWELL, T., HOLROYD, S., *et al*. 2000. The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**: 665-668.

PASSALACQUA, K.D., VARADARAJAN, A., ONDOV, B.D., OKOU, D.T., ZWICK, M.E. & BERGMAN, N.H. 2009. Structure and complexity of a bacterial transcriptome. *Journal of Bacteriology* **191**: 3203-3211.

PEARSON, W.R. 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods in Molecular Biology* **132**: 185-219.

PEDDIREDDI, L., CHENG, C. & GANTA, R.R. 2009. Promoter analysis of macrophage- and tick cell-specific differentially expressed *Ehrlichia chaffeensis* p28-Omp genes. *BMC Microbiology* **9**: 99.

PERKINS, T.T., KINGSLEY, R.A., FOOKES, M.C., GARDNER, P.P., JAMES, K.D., YU, L., ASSEFA, S.A., HE, M., CROUCHER, N.J., PICKARD, D.J., *et al*. 2009. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella* Typhi. *PLoS Genetics* **5**: e1000569.

PETES, T.D. & HILL, C.W. 1988. Recombination between repeated genes in microorganisms. *Annual Review of Genetics* **22**: 147-168.

PETNEY, T.N., HORAK, I.G. & RECHAV, Y. 1987. The ecology of the African vectors of heartwater, with particular reference to *Amblyomma hebraeum* and *Amblyomma variegatum*. *Onderstepoort Journal of Veterinary Research* **54**: 381-395.

PIZZA, M., SCARLATO, V., MASIGNANI, V., GIULIANI, M.M., ARICÒ, B., COMANDUCCI, M., JENNINGS, G.T., BALDI, L., BARTOLINI, E., CAPECCHI, B., *et al*. 2000. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **287**: 1816-1820.

POP, M. 2009. Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics* **10**: 354-366.

POSTIGO, M., TAOUFIK, A., BELL-SAKYI, L., DE VRIES, E., MORRISON, W.I. & JONGEJAN, F. 2007. Differential transcription of the major antigenic protein 1 multigene family of *Ehrlichia ruminantium* in *Amblyomma variegatum* ticks. *Veterinary Microbiology* **122**: 298-305.

POWELL, K. 2004. DNA vaccines – back in the saddle again? *Nature Biotechnology* **22**: 799-801.

PRESTON, P.M., BROWN, C.G.D. & RICHARDSON, W. 1992. Cytokines inhibit the development of trophozoite-infected cells of *Theileria annulata* and *Theileria parva* but enhance the proliferation of macroschizont-infected cell lines. *Parasite Immunology* **14**: 125-141.

PRETORIUS, A., COLLINS, N.E., STEYN, H.C. & ALLSOPP, B.A. 2002a. Sequence analysis of three *Ehrlichia ruminantium* LambdaGEM-11 clones. *Annals of the New York Academy of Sciences* **969**: 155-158.

PRETORIUS, A., VAN STRIJP, F., BRAYTON, K.A., COLLINS, N.E. & ALLSOPP, B.A. 2002b. Genetic immunization with *Ehrlichia ruminantium* GroEL and GroES homologues. *Annals of the New York Academy of Sciences* **969**: 151-154.

PRETORIUS, A., COLLINS, N.E., STEYN, H.C., VAN STRIJP, F., VAN KLEEF, M. & ALLSOPP, B.A. 2007. Protection against heartwater by DNA immunisation with four *Ehrlichia ruminantium* open reading frames. *Vaccine* **25**: 2316-2324.

PRETORIUS, A., VAN KLEEF, M., COLLINS, N.E., TSHIKUDO, N., LOUW, E., FABER, F.E., VAN STRIJP, M.F. & ALLSOPP B.A. 2008. A heterologous prime/boost immunisation strategy protects against virulent *E. ruminantium* Welgevonden needle challenge but not against tick challenge. *Vaccine* **26**: 4363-4371.

PRETORIUS, A., LIEBENBERG, J., LOUW, E., COLLINS, N.E. & ALLSOPP B.A. 2010. Studies of a polymorphic *E. ruminantium* gene for use as a component of a recombinant vaccine against heartwater. *Vaccine* **28**: 3531-3539.

PROVOST, A. & BEZUIDENHOUT, J.D. 1987. The historical background and global importance of heartwater. *Onderstepoort Journal of Veterinary Research* **54**: 165-169.

PROZESKY, L. & DU PLESSIS, J.L. 1987. Heartwater. The development and life cycle of *Cowdria ruminantium* in the vertebrate host, ticks and cultured endothelial cells. *Onderstepoort Journal of Veterinary Research* **54**: 193-196.

QI, W., KÄSER, M., RÖLTGEN, K., YEBOAH-MANU, D. & PLUSCHKE, G. 2009. Genomic diversity and evolution of *Mycobacterium ulcerans* revealed by next-generation sequencing. *PLoS Pathogens* **5**: e1000580.

RAFATI, S., ZAHEDIFARD, F. & NAZGOUEE, F. 2006. Prime-boost vaccination using cysteine proteinases type I and II of *Leishmania infantum* confers protective immunity in murine visceral leishmaniasis. *Vaccine* **24**: 2169-2175.

RAMARAO, N., GRAY-OWEN, S.D., BACKERT, S. & MEYER, T.F. 2000. *Helicobacter pylori* inhibits phagocytosis by professional phagocytes involving type IV secretion components. *Molecular Microbiology* **37**: 1389-1404.

RAPP, U.K. & KAUFMANN, S.H.E. 2004. DNA vaccination with gp96-peptide fusion proteins induces protection against an intracellular bacterial pathogen. *International Immunology* **16**: 597-605.

RAPPUOLI, R. 2000. Reverse vaccinology. *Current Opinion in Microbiology* **3**: 445-450.

RAPPUOLI, R. 2001. Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine* **19**: 2688-2691.

RAPPUOLI, R. 2007. Bridging the knowledge gaps in vaccine design. *Nature Biotechnology* **25**: 1361-1366.

REDDY, G.R. 1995. Determining the sequence of parasite DNA. *Parasitology Today* **11**: 37-42.

RILEY, M. 1993. Functions of the gene products of *Escherichia coli*. *Microbiological Reviews* **57**: 862-952.

ROCHA, E.P. 2003. An appraisal of the potential for illegitimate recombination in bacterial genomes and its consequences: from duplications to genome reduction. *Genome Research* **13**: 1123-1132.

ROITT, I. 1991. Essential immunology. 8th ed. Oxford: Blackwell Scientific Publications, pp 3-20.

RONAGHI, M. 2001. Pyrosequencing sheds light on DNA sequencing. *Genome Research* **11**: 3-11.

ROSET, M.S., CIOCCHINI, A.E., UGALDE, R.A. & IÑÓN DE IANNINO, N. 2004. Molecular cloning and characterization of *cgt*, the *Brucella abortus* cyclic beta-1,2-glucan transporter gene, and its role in virulence. *Infection and Immunity* **72**: 2263-2271.

ROSS, B.C., CZAJKOWSKI, L., HOCKING, D., MARGETTS, M., WEBB, E., ROTHEL, L., PATTERSON, M., AGIUS, C., CAMUGLIA, S., REYNOLDS, E., *et al.* 2001. Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis*. *Vaccine* **19**: 4135-4142.

ROSSOUW, M., NEITZ, A.W., DE WAAL, D.T., DU PLESSIS, J.L., VAN GAS, L. & BRETT, S. 1990. Identification of the antigenic proteins of *Cowdria ruminantium*. *Onderstepoort Journal of Veterinary Research* **57**: 215-221.

RUTHERFORD, K., PARKHILL, J., CROOK, J., HORSNELL, T., RICE, P., RAJANDREAM, M.-A. & BARRELL, B. 2000. Artemis: sequence visualisation and annotation. *Bioinformatics* **16**: 944-945.

SAITOU, N. & NEI, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* **4**: 406-425.

SAMBROOK, J., FRITSCH, E.H. & MANIATIS, T. 1989. Molecular Cloning: A Laboratory Manual, 2nd ed. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y., pp E3-E4.

SANGER, F., NICKLEN, S. & COULSON, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America* **74**: 5463-5467.

SEDEGAH, M., CHAROENVIT, Y., MINH, L., BELMONTE, M., MAJAM, V.F., ABOT, S., GANESHAN, H., KUMAR, S., BACON, D.J., STOWERS, A., *et al.* 2004. Reduced immunogenicity of DNA vaccine plasmids in mixtures. *Gene Therapy* **11**: 448-456.

SEGAL, E.D., CHA, J., LO, J., FALKOW, S. & TOMPKINS, L.S. 1999. Altered states: involvement of phosphorylated CagA in the induction of host cellular growth changes by *Helicobacter pylori*. *Proceedings of the National Academy of Sciences of the United States of America* **96**: 14559-14564.

SERRUTO, D. & RAPPUOLI, R. 2006. Post-genomic vaccine development. *FEBS Letters* **580**: 2985-2992.

SERRUTO, D., SERINO, L., MASIGNANI, V. & PIZZA, M. 2009. Genome-based approaches to develop vaccines against bacterial pathogens. *Vaccine* **27**: 3245-3250.

SHE, Q., CONFALONIERI, F., ZIVANOVIC, Y., MEDINA, N., BILLAULT, A., AWAYEZ, M.J., THI-NGOC, H.P., PHAM, B.T., VAN DER OOST, J., DUGUET, M. & GARRETT, R.A. 2000. A BAC library and paired-PCR approach to mapping and completing the genome sequence of *Sulfolobus solfataricus* P2. *DNA Sequence* **11**: 183-192.

SHENDURE, J., PORRECA, G.J., REPPAS, N.B., LIN, X., MCCUTCHEON, J.P., ROSENBAUM, A.M., WANG, M.D., ZHANG, K., MITRA, R.D. & CHURCH, G.M. 2005. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**: 1728-1732.

SIGRIST, C.J.A., CERUTTI, L., HULO, N., GATTIKER, A., FALQUET, L., PAGNI, M., BAIROCH, A. & BUCHER, P. 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics* **3**: 265-274.

SIMPSON, B.C., LINDSAY, M.S., MORRIS, J.R., MUIRHEAD, F.S., POLLOCK, A., PRICHARD, S.G., STANLEY, H.G., THIRLWELL, G.R., HUNTER, A.G., BRADLEY, J., *et al.* 1987. Protection of cattle against heartwater in Botswana: comparative efficacy of different methods against natural and blood-derived challenges. *The Veterinary Record* **120**: 135-138.

SINGU, V., LIU, H., CHENG, C. & GANTA, R.R. 2005. *Ehrlichia chaffeensis* expresses macrophage- and tick cell-specific 28-kilodalton outer membrane proteins. *Infection and Immunity* **73**: 79-87.

SINGU, V., PEDDIREDDI, L., SIRIGIREDDY, K.R., CHENG, C., MUNDERLOH, U. & GANTA, R.R. 2006. Unique macrophage and tick cell-specific protein expression from the p28/p30-outer membrane protein multigene locus in *Ehrlichia chaffeensis* and *Ehrlichia canis*. *Cellular Microbiology* **8**: 1475-1487.

SONNHAMMER, E.L. & DURBIN, R. 1994. A workbench for large-scale sequence homology analysis. *Computer Applications in the Biosciences* **10**: 301-307.

SONNHAMMER, E.L. & DURBIN, R. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167**: GC1-10.

SPRENGER, J., FINK, J.L. & TEASDALE, R.D. 2006. Evaluation and comparison of mammalian subcellular localization prediction methods. *BMC Bioinformatics* **7**: S3.

SPREULL, J. 1904. Heartwater inoculation experiments. *Agricultural Journal of the Cape of Good Hope* **24**: 433-442.

STADEN, R., BEAL, K.F. & BONFIELD, J.K. 2000. The Staden package, 1998. *Methods in Molecular Biology* **132**: 115-130.

STEMKE-HALE, K., KALTENBOECK, B., DEGRAVES, F.J., SYKES, K.F., HUANG, J., BU, C.H. & JOHNSTON, S.A. 2005. Screening the whole genome of a pathogen *in vivo* for individual protective antigens. *Vaccine* **23**: 3016-3025.

STEPHENS, R.S., KALMAN, S., LAMMEL, C., FAN, J., MARATHE, R., ARAVIND, L., MITCHELL, W., OLINGER, L., TATUSOV, R.L., ZHAO, Q., *et al*. 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. *Science* **282**: 754-759.

STEYN, H.C., PRETORIUS, A., MCCRINDLE, C.M.E., STEINMANN, C.M.L. & VAN KLEEF, M. 2008. A quantitative real-time PCR assay for *Ehrlichia ruminantium* using pCS20. *Veterinary Microbiology* **131**: 258-265.

STORM, A.J., STORM, C., CHEN, J., ZANDBERGEN, H., JOANNY, J.F. & DEKKER, C. 2005. Fast DNA translocation through a solid-state nanopore. *Nano Letters* **5**: 1193-1197.

STOTHARD, P. 2000. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *BioTechniques* **28**: 1102-1104.

STRAUSS, E.C., KOBORI, J.A., SIU, G. & HOOD, L.E. 1986. Specific-primer-directed DNA sequencing. *Analytical Biochemistry* **154**: 353-360.

SUNDQUIST, A., RONAGHI, M., TANG, H., PEVZNER, P. & BATZOGLOU, S. 2007. Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS ONE* **2**: e484.

SULSONA, C.R., MAHAN, S.M. & BARBET, A.F. 1999. The *map*1 gene of *Cowdria ruminantium* is a member of a multigene family containing both conserved and variable genes. *Biochemical and Biophysical Research Communications* **257**: 300-305.

SYKES, K.F. & JOHNSTON, S.A. 1999. Genetic live vaccines mimic the antigenicity but not pathogenicity of live viruses. *DNA and Cell Biology* **18**: 521-531.

TANG, F., BARBACIORU, C., WANG, Y., NORDMAN, E., LEE, C., XU, N., WANG, X., BODEAU, J., TUCH, B.B., SIDDIQUI, A., *et al*. 2009. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**: 377-382.

TETTELIN, H., SAUNDERS, N.J., HEIDELBERG, J., JEFFRIES, A.C., NELSON, K.E., EISEN, J.A., KETCHUM, K.A., HOOD, D.W., PEDEN, J.F., DODSON, R.J., *et al*. 2000. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* **287**: 1809-1815.

The Gene Ontology Consortium. 2000. Gene Ontology: Tool for the unification of biology. *Nature Genetics* **25**: 25-29.

THOMPSON, D.V., MELCHERS, L.S., IDLER, K.B., SCHILPEROORT, R.A. & HOOYKAAS, P.J. 1988. Analysis of the complete nucleotide sequence of the *Agrobacterium tumefaciens vir*B operon. *Nucleic Acids Research* **16**: 4621-4636.

THOMPSON, J.D., GIBSON, T.J. & HIGGINS, D.G. 2002. Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics* **2**: Unit 2.3.

TOTTÉ, P., BLANKAERT, D., ZILIMWABAGABO, P. & WERENNE, J. 1993. Inhibition of *Cowdria ruminantium* infectious yield by interferons alpha and gamma in endothelial cells. *Revue d'Élevage et de Médecine Vétérinaire des Pays Tropicaux* **46**: 189-194.

TOTTÉ, P., VACHIERY, N., MARTINEZ, D., TRAP, I., BALLINGALL, K.T., MACHUGH, N.D., BENSAID, A. & WERENNE, J. 1996. Recombinant bovine interferon gamma inhibits the growth of *Cowdria ruminantium* but fails to induce major histocompatibility complex class II following infection of endothelial cells. *Veterinary Immunology and Immunopathology* **53**: 61-71.

TOTTÉ, P., MCKEEVER, D., MARTINEZ, D. & BENSAID, A. 1997. Analysis of T-cell responses in cattle immunized against heartwater by vaccination with killed elementary bodies of *Cowdria ruminantium*. *Infection and Immunity* **65**: 236-241.

TOTTÉ, P., BENSAID, A., MAHAN, S.M., MARTINEZ, D. & MCKEEVER, D.J. 1999. Immune responses to *Cowdria ruminantium* infections. *Parasitology Today* **15**: 286-290.

TURNER, D.J., KEANE, T.M., SUDBERY, I. & ADAMS, D.J. 2009. Next-generation sequencing of vertebrate experimental organisms. *Mammalian Genome* **20**: 327-338.

UILENBERG, G. 1982. Experimental transmission of *Cowdria ruminantium* by the Gulf coast tick *Amblyomma maculatum*: danger of introducing heartwater and benign African Theileriasis onto the American mainland. *American Journal of Veterinary Research* **43**: 1279-1282.

UILENBERG, G. 1983. Heartwater (*Cowdria ruminantium* infection): current status. *Advances in Veterinary Science and Comparative Medicine* **27**: 427-480.

UILENBERG, G. 1990. Extension de la tique *Amblyomma variegatum* dans les Antilles: comment expliquer cette grave menace et que faire? *Revue d'Élevage et de Médecine Vétérinaire des Pays Tropicaux* **43**: 297-299.

ULMER, J.B., WAHREN, B. & LIU, M.A. 2006. Gene-based vaccines: recent technical and clinical advances. *Trends in Molecular Medicine* **12**: 216-222.

VACHIÉRY, N., LEFRANÇOIS, T., ESTEVES, I., MOLIA, S., SHEIKBOUDOU, C., KANDASSAMY, Y. & MARTINEZ, D. 2006. Optimisation of the inactivated vaccine dose against heartwater and in vitro quantification of *Ehrlichia ruminantium* challenge material. *Vaccine* **24**: 4747-4756.

VAN BELKUM, A., SCHERER, S., VAN ALPHEN, L. & VERBRUGH, H. 1998. Short-sequence DNA repeats in prokaryotic genomes. *Microbiology and Molecular Biology Reviews* **62**: 275-293.

VANBUSKIRK, K.M., O'NEILL, M.T., DE LA VEGA, P., MAIER, A.G., KRZYCH, U., WILLIAMS, J., DOWLER, M.G., SACCI, J.B. Jr., KANGWANRANGSAN, N., TSUBOI, T., *et al.* 2009. Preerythrocytic, live-attenuated *Plasmodium falciparum* vaccine candidates by design. *Proceedings of the National Academy of Sciences of the United States of America* **106**: 13004-13009.

VAN DE PYPEKAMP, H.E. & PROZESKY, L. 1987. Heartwater. An overview of the clinical signs, susceptibility and differential diagnoses of the disease in domestic ruminants. *Onderstepoort Journal of Veterinary Research* **54**: 263-266.

VAN HEERDEN, H., COLLINS, N.E., ALLSOPP, M.T.E.P. & ALLSOPP, B.A. 2002. Major outer membrane proteins of *Ehrlichia ruminantium* encoded by multigene family. *Annals of the New York Academy of Sciences* **969**: 131-134.

VAN HEERDEN, H., COLLINS, N.E., BRAYTON, K.A., RADEMEYER, C. & ALLSOPP, B.A. 2004a. Characterization of a major outer membrane protein multigene family in *Ehrlichia ruminantium*. *Gene* **330**: 159-168.

VAN HEERDEN, H., STEYN, H.C., ALLSOPP, M.T.E.P., ZWEYGARTH, E., JOSEMANS, A.I. & ALLSOPP, B.A. 2004b. Characterization of the pCS20 region of different *Ehrlichia ruminantium* isolates. *Veterinary Microbiology* **101**: 279-291.

VAN KLEEF, M., GUNTER, N.J., MACMILLAN, H., ALLSOPP, B.A., SHKAP, V. & BROWN, W.C. 2000. Identification of *Cowdria ruminantium* antigens that stimulate proliferation of lymphocytes from cattle immunized by infection and treatment or with inactivated organisms. *Infection and Immunity* **68**: 603-614.

VAN KLEEF, M., MACMILLAN, H., GUNTER, N.J., ZWEYGARTH, E., ALLSOPP, B.A., SHKAP, V., DU PLESSIS, D.H. & BROWN, W.C. 2002. Low molecular weight proteins of *Cowdria ruminantium* (Welgevonden isolate) induce bovine CD4+-enriched T-cells to proliferate and produce interferon-γ. *Veterinary Microbiology* **85**: 259-273.

VAN VLIET, A.H., JONGEJAN, F., VAN KLEEF, M. & VAN DER ZEIJST, B.A. 1994. Molecular cloning, sequence analysis, and expression of the gene encoding the immunodominant 32-kilodalton protein of *Cowdria ruminantium*. *Infection and Immunity* **62**: 1451-1456.

VAN VLIET, A.H. 2010. Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS Microbiology Letters* **302**: 1-7.

VENTER, J.C., ADAMS, M.D., MYERS, E.W., LI, P.W., MURAL, R.J., SUTTON, G.G., SMITH, H.O., YANDELL, M., EVANS, C.A., HOLT, R.A., *et al.* 2001. The sequence of the human genome. *Science* **291**: 1304-1351.

VOELKERDING, K.V., DAMES, S.A. & DURTSCHI, J.D. 2009. Next-generation sequencing: from basic research to diagnostics. *Clinical Chemistry* **55**: 641-658.

VON HEIJNE, G. 1985. Signal sequences. The limits of variation. *Journal of Molecular Biology* **184**: 99-105.

VON HEIJNE, G. 1992. Membrane protein structure prediction – Hydrophobicity analysis and the positive-inside rule. *Journal of Molecular Biology* **225**: 487-494.

VOYICH, J.M., STURDEVANT, D.E., BRAUGHTON, K.R., KOBAYASHI, S.D., LEI, B., VIRTANEVA, K., DORWARD, D.W., MUSSER, J.M. & DELEO, F.R. 2003. Genome-wide protective response used by group A *Streptococcus* to evade destruction by human polymorphonuclear leukocytes. *Proceedings of the National Academy of Sciences of the United States of America* **100**: 1996-2001.

WAGHELA, S.D., RURANGIRWA, F.R., MAHAN, S.M., YUNKER, C.E., CRAWFORD, T.B., BARBET, A.F., BURRIDGE, M.J. & MCGUIRE, T.C. 1991. A cloned DNA probe identifies *Cowdria ruminantium* in *Amblyomma variegatum* ticks. *Journal of Clinical Microbiology* **29**: 2571-2577.

WAKEEL, A., KURIAKOSE, J.A. & MCBRIDE, J.W. 2009. An *Ehrlichia chaffeensis* tandem repeat protein interacts with multiple host targets involved in cell signaling, transcriptional regulation, and vesicle trafficking. *Infection and Immunity* **77**: 1734-1745.

WALKER, J.B. & OLWAGE, A. 1987. The tick vectors of *Cowdria ruminantium* (Ixodoidea, Ixodiae, genus *Amblyomma*) and their distribution. *Onderstepoort Journal of Veterinary Research* **54**: 353-379.

WANG, Q.M., SUN, S.H., HU, Z.L., YIN, M., XIAO, C.J. & ZHANG, J.C. 2004a. Improved immunogenicity of a tuberculosis DNA vaccine encoding ESAT6 by DNA priming and protein boosting. *Vaccine* **22**: 3622-3627.

WANG, R., EPSTEIN, J., CHAROENVIT, Y., BARACEROS, F.M., RAHARDJO, N., GAY, T., BANANIA, J.G., CHATTOPADHYAY, R., DE LA VEGA, P., RICHIE, T.L., *et al*. 2004b. Induction in humans of CD8+ and CD4+ T cell and antibody responses by sequential immunization with malaria DNA and recombinant protein. *The Journal of Immunology* **172**: 5561-5569.

WANG, J., WANG, W., LI, R., LI, Y., TIAN, G., GOODMAN, L., FAN, W., ZHANG, J., LI, J., ZHANG, J., *et al*. 2008. The diploid genome sequence of an Asian individual. *Nature* **456**: 60-65.

WANG, Z., GERSTEIN, M. & SNYDER, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics* **10**: 57-63.

WANG, T. & YANG, J. 2009. Using the nonlinear dimensionality reduction method for the prediction of subcellular localization of Gram-negative bacterial proteins. *Molecular Diversity* **13**: 475-481.

WARREN, R.L., SUTTON, G.G., JONES, S.J. & HOLT, R.A. 2006. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics* **23**: 500-501.

WEAVER, R.F. & HEDRICK, P.W. 1992. Genetics. 2nd ed. Dubuque: Wm. C. Brown Publishers, pp 296-302.

WEISS, K.E., HAIG, D.A. & ALEXANDER, R.A. 1952. Aureomycin in the treatment of heartwater. *Onderstepoort Journal of Veterinary Research* **25**: 41-50.

WEISS, A.A., JOHNSON, F.D. & BURNS, D.L. 1993. Molecular characterization of an operon required for pertussis toxin secretion. *Proceedings of the National Academy of Sciences of the United States of America* **90**: 2970-2974.

WEITZMANN, M.N., WOODFORD, K.J. & USDIN, K. 1997. DNA secondary structures and the evolution of hypervariable tandem arrays. *Journal of Biological Chemistry* **272**: 9517-9523.

WHEELER, D.A., SRINIVASAN, M., EGHOLM, M., SHEN, Y., CHEN, L., MCGUIRE, A., HE, W., CHEN, Y.J., MAKHIJANI, V., ROTH, G.T., *et al*. 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**: 872-876.

WHITEFORD, N., HASLAM, N., WEBER, G., PRÜGEL-BENNETT, A., ESSEX, J.W., ROACH, P.L., BRADLEY, M. & NEYLON, C. 2005. An analysis of the feasibility of short read sequencing. *Nucleic Acids Research* **33**: e171.

WICKER, T., SCHLAGENHAUF, E., GRANER, A., CLOSE, T.J., KELLER, B. & STEIN, N. 2006. 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* **7**: 275.

WILLIAMS, K.P., SOBRAL, B.W. & DICKERMAN, A.W. 2007. A robust species tree for the Alphaproteobacteria. *Journal of Bacteriology* **189**: 4578-4586.

WILSON, C.G., KAJANDER, T. & REGAN, L. 2005. The crystal structure of NlpI. A prokaryotic tetratricopeptide repeat protein with a globular fold. *FEBS Journal* **272**: 166-179.

WIUF, C. & HEIN, J. 2000. The coalescent with gene conversion. *Genetics* **155**: 451-462.

WIZEMANN, T.M., HEINRICHS, J.H., ADAMOU, J.E., ERWIN, A.L., KUNSCH, C., CHOI, G.H., BARASH, S.C., ROSEN, C.A., MASURE, H.R., TUOMANEN, E., *et al*. 2001. Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. *Infection and Immunity* **69**: 1593-1598.

WOOSTER, R., BIGNELL, G., LANCASTER, J., SWIFT, S., SEAL, S., MANGION, J., COLLINS, N., GREGORY, S., GUMBS, C. & MICKLEM, G. 1995. Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**: 789-792.

WU, M., SUN, L.V., VAMATHEVAN, J., RIEGLER, M., DEBOY, R., BROWNLIE, J.C., MCGRAW, E.A., MARTIN, W., ESSER, C., AHMADINEJAD, N., *et al*. 2004. Phylogenomics of the reproductive parasite *Wolbachia pipientis w*Mel: a streamlined genome overrun by mobile genetic elements. *PLoS Biology* **2**: 0327-0341.

XIA, Q., GUO, Y., ZHANG, Z., LI, D., XUAN, Z., LI, Z., DAI, F., LI, Y., CHENG, D., LI, R., *et al*. 2009. Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (Bombyx). *Science* **326**: 433-436.

YANG, X., HUDSON, M., WALTERS, N., BARGATZE, R.F. & PASCUAL, D.W. 2005. Selection of protective epitopes for *Brucella melitensis* by DNA vaccination. *Infection and Immunity* **73**: 7297-7303.

YU, C.S., LIN, C.J. & HWANG, J.K. 2004. Predicting subcellular localisation of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. *Protein Science* **13**: 1402-1406.

YU, N.Y., WAGNER, J.R., LAIRD, M.R., MELLI, G., REY, S., LO, R., DAO, P., SAHINALP, S.C., ESTER, M., FOSTER, L.J. & BRINKMAN, F.S. 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26**: 1608-1615.

YUAN, Z., DAVIS, M.J., ZHANG, F. & TEASDALE, R.D. 2003. Computational differentiation of N-terminal signal peptides and transmembrane helices. *Biochemical and Biophysical Research Communications* **312**: 1278-1283.

ZARLENGA, D.S. & GASBARRE, L.C. 2009. From parasite genomes to one healthy world: Are we having fun yet? *Veterinary Parasitology* **163**: 235-249.

ZHANG, X., LUO, T., KEYSARY, A., BANETH, G., MIYASHIRO, S., STRENGER, C., WANER, T. & MCBRIDE, J.W. 2008a. Genetic and antigenic diversities of major immunoreactive proteins in globally distributed *Ehrlichia canis* strains. *Clinical and Vaccine Immunology* **15**: 1080-1088.

ZHANG, C., XIONG, Q., KIKUCHI, T. & RIKIHISA, Y. 2008b. Identification of 19 polymorphic major outer membrane protein genes and their immunogenic peptides in *Ehrlichia ewingii* for use in a serodiagnostic assay. *Clinical and Vaccine Immunology* **15**: 402-411.

ZHU, B., NETHERY, K.A., KURIAKOSE, J.A., WAKEEL, A., ZHANG, X. & MCBRIDE, J.W. 2009. Nuclear translocated *Ehrlichia chaffeensis* ankyrin protein interacts with a specific adenine-rich motif of host promoter and intronic Alu elements. *Infection and Immunity* **77**: 4243-4255.

ZIENTZ, E., DANDEKAR, T. & GROSS, R. 2004. Metabolic interdependence of obligate intracellular bacteria and their insect hosts. *Microbiology and Molecular Biology Reviews* **68**: 745-770.

ZIVKOVIC, Z., ESTEVES, E., ALMAZÁN, C., DAFFRE, S., NIJHOF, A.M., KOCAN, K.M., JONGEJAN, F. & DE LA FUENTE, J. 2010. Differential expression of genes in salivary glands of male *Rhipicephalus (Boophilus)microplus* in response to infection with *Anaplasma marginale*. *BMC Genomics* **11**: 186.

ZWEYGARTH, E. & JOSEMANS, A.I. 2001a. A chemically defined medium for the growth of *Cowdria ruminantium*. *Onderstepoort Journal of Veterinary Research* **68**: 37-40.

ZWEYGARTH, E. & JOSEMANS, A.I. 2001b. Continuous *in vitro* propagation of *Cowdria ruminantium* (Welgevonden stock) in a canine macrophage-monocyte cell line. *Onderstepoort Journal of Veterinary Research* **68**: 155-157.

ZWEYGARTH, E., JOSEMANS, A.I., VAN STRIJP, M.F., LOPEZ-REBOLLAR, L., VAN KLEEF, M. & ALLSOPP, B.A. 2005. An attenuated *Ehrlichia ruminantium* (Welgevonden stock) vaccine protects small ruminants against virulent heartwater challenge. *Vaccine* **23**: 1695-1702.

ZWEYGARTH, E., JOSEMANS, A.I. & STEYN, H.C. 2008. Experimental use of the attenuated *Ehrlichia ruminantium* (Welgevonden) vaccine in Merino sheep and Angora goats. *Vaccine* **26S**: G34-39.

Appendix B: Materials, buffers, media and solutions

B1: Suppliers of the materials used in this study

Supplier	Product
Amersham Biosciences	pMOS*Blue* blunt ended cloning kit, TempliPhi DNA Sequencing Template Amplification Kit, [^{3}H] thymidine
Applied Biosystems	Amplitaq Gold polymerase, Dye-terminator Cycle Sequencing kit
BDH	Glycine, glycerol, magnesium chloride, potassium bi-phosphate
Bio-Rad	Biolistic® 1.6 Micron Gold, Criterion™ XT precast gels, RC/DC Protein Assay, XT MOPS Running Buffer, XT Sample Buffer, XT Reducing Agent, Bio-Safe Coomassie
Celtic Molecular Diagnostics	mouse anti-bovine IFN-γ mAb CC302
Centaur	Eutha-Nase
Costar	Half-area 96-well plates
Gibco BRL Products	Concert Rapid PCR Purification kit
Highveldt biologicals	Foetal calf serum
Immonodiagnostik	rabbit anti-bovine IFN-γ anti-serum
Inqaba Biotec	Primers
Invitek	Invisorb® Spin Plasmid Mini *Two* kit, MSB® Spin PCRapace kit
Invitrogen	pET102/D-TOPO® expression system, Platinum® *pfx* DNA polymerase kit, RPMI-1640 + GlutaMAX-I
Macherey-Nagel	NucleoBond® Xtra Maxi purification Kit, Protino®Ni 1000 prepacked columns kit
Merck Biosciences	Acetone, calcium chloride, chloroform, citric acid, ethanol, glucose, isoamyl alcohol, methanol, potassium chloride, sodium carbonate, sodium mono-phosphate, tri-sodium citrate, tryptone, Tween-20
Millipore Corporation	dH2O, PVDF membranes, MAIPS 4510 Multiscreen™-IP filtration plates
Montanide	ISA50
MP Biomedicals	Penicillin, streptomycin
Novagen	Overnight Express Autoinduction system 1, BugBuster® Protein Extraction Reagent
Ondestepoort Biological Products	Elsevers medium, PBS
Packard BioScience	Ultima gold F scintillation solution
Pfizer	Liquamycin/LA
Pierce	SuperSignal®West Pico Chemiluminescent substrate
Promega	Ethidium bromide solution, isopropyl-β-D-thiogalactopyranoside (IPTG), *Pfu* polymerase, pGEM-T Easy cloning kit, shrimp alkaline phosphatase, T4 ligase
Qiagen	QIAquick PCR Purification Kit, QIAprep 8 Miniprep Kit,
Roche	Anti-His_6 antibodies, DNAse I, High Pure PCR product purification kit, High Pure Plasmid Isolation kit, Restriction endonucleases (*Bam*HI, *Eco*RI, *Sal*I and *Xba*I)
Sigma	Ampicillin, anti-rabbit IgG alkaline phosphatase conjugate, bromophenol blue, concanavalin A, Dulbecco's PBS, Fast BCIP/NBT substrate tablets, Hank's Balanced Salt Solution, Histopaque®-1077, 2-mercaptoethanol, percoll, phenol, proteinase K, RNase, spermidine, trypan blue

Supplier	Product
Stratagene	Klenow Fill-In kit
TaKaRa Bio Inc.	TaKaRa Ex Taq, TaKaRa recochips
USB	Agar, boric acid, EDTA, HCl, magnesium sulphate, potassium bi-phosphate, SDS, sodium chloride, Tris base
Walac	96 well glass fibre filters
White Sci	Agarose
Zymed	HRP-goat-anti-mouse IgG

B2: Preparation of buffers, media and solutions

B2.1. Buffers for general laboratory use

Ampicillin

Prepare a stock solution of 10 µg/ml by dissolving 50 mg ampicillin in 5 ml dH_2O. The solution was filter sterilised and stored in aliquots at -20ºC.

1M IPTG

Dissolve 1.19 g isopropyl-β-D-thiogalactopyranoside in 50 ml dH_2O. Sterilise by filtration, aliquot and store at -20ºC.

LB Agar plates

Dissolve 10 g tryptone, 5 g yeast, 10 g NaCl and 15 g bacto-agar in 800 ml dH_2O. Adjust volume to 1000 ml with dH_2O and sterilise by autoclaving. Allow to cool to 55ºC and add the appropriate amount of antibiotics before pouring the plates. Store at 4ºC.

LB broth

Dissolve 10 g tryptone, 5 g yeast and 10 g NaCl in 800 ml dH_2O. Adjust volume to 1000 ml with dH_2O and sterilise by autoclaving. Store at 4ºC and add appropriate antibiotics prior to use.

SOC

Buffer consists of 20 g tryptone, 5 g yeast extract, 0.5 g NaCl and 2.5 ml 1 M KCl. Adjust pH to 7 with NaOH, make up to 970 ml and autoclave. Add sterile 1 M MgCl2 and 20 ml sterile glucose prior to use.

10x TBE

Dissolve 108 g Tris base, 55 g boric acid and 7.44 g EDTA in 800 ml dH_2O. Adjust volume to 1000 ml with dH_2O.

X-gal

Dissolve 400 mg 5-bromo-4-chloro-3-indolyl-β-D-galactoside (X-gal) in 20 ml N,N'-dimethyl formamide. Aliquot and store at -20ºC.

B2.2. Western blots buffers

Blocking buffer

Dissolve 1 g bovine serum albumin in 100 ml 1x PBS.

Transfer buffer

Dissolve 9.1 g Tris base (38 mM), 43.25 g glycine (288 mM) in 1700 ml dH_2O and add 300 ml methanol (pH ~8.3-8.4).

Wash buffer

Dissolve 500 µl Tween-20 in 1000 ml 1x PBS.

B2.3. ELISpot buffers

Blocking medium

RPMI-1640 medium supplemented with 10% heat inactivated foetal calf serum.

Coating antibody

Dissolve mouse anti-bovine IFN-γ mAb CC302 in sterile PBS (100 µg/ml), aliquot and store at -20ºC.

Carbonate/bicarbonate coating buffer

Buffer consists of 15mM Na_2CO_3 and 35 mM $NaHCO_3$. Adjust pH to 9.6, and filter sterilize.

Complete RPMI-1640 medium

Add to RPMI-1640 + GlutaMAX-l: 10% foetal calf serum, 5 x 10^{-5} M mercaptoethanol, 100 U/ml penicillin and 0.1 mg/ml streptomycin.

dH_2O-T

dH_2O with 0.05% Tween-20.

PBS-T

PBS with 0.05% Tween-20.

PBS-T/BSA

PBS-T with 0.1% bovine serum albumin

Rabbit anti-bovine IFN-γ antiserum

Dissolve rabbit anti bovine IFN- γ antibody in 100 µl dH_2O and sterilise through a 0.22 µm filter.

Monoclonal anti-rabbit IgG alkaline phosphate conjugate

Dissolve 1 tablet/10 ml dH_2O at room temperature for 30 min and sterilise through a 0.22 µm filter.

Appendix C: Primers

C1: Primers used to complete the genome sequence.

Primer name	Position	Primer sequence (5' > 3')
WTHIN992_1F	107 > 127	GGATTAGGACGTATAGGAAGG
WTHIN992_1R	561 > 539	CAGTTTGTTGTACATGATCCAGC
WTHIN2864_27R	814 > 793	GGAAACATTTGGTGTTGGTACG
WGAP49_2F	1050 > 1072	TGAGAATGCTGGATATAGTGTGC
WGAP49_F	1315 > 1350	TTTTTACAATGATTTATATATAACCCTATACAACAG
WGAP49walk_1F	1556 > 1583	ACAGAAAAGTAACACCTAATAACAATGC
WGAP49walk_1R	2467 > 2436	TGTTTGTTGGGTATTGTAGTTAATTATATAGC
WGAP49_2R	2858 > 2835	TTGTCCAACCATAAAATCTGTACG
1740_R	3321 > 3298	TTAGAGAAGCTGTGTTAGTTATGC
WGAP49_R	3331 > 3308	GGTTCAGAAATTAGAGAAGCTGTG
WTHIN24731_1F	5395 > 5420	CCAGATGTAATAGAAATAGATGCAGC
WTHIN24731_1R	6977 > 6956	CTCTCCCTACATTTCCTGTAGC
WTHIN24731_2F	8171 > 8197	TCACATGAAACATAACATTCAATTCCC
WTHIN24731_2R	9741 > 9717	AGATCTTGCTTTGCAAATAAATCGC
WTHIN24731_3F	12680 > 12707	GGCATTATTTACAACATGTATTAATGGC
WTHIN24731_3R	13900 > 13874	CTTGTTAATGATGGTAGATCTATACGC
WTHIN24731_4R	15023 > 15003	CACCAATGCTTGGATTTTCCC
WTHIN24731_5F	22812 > 22837	GCATCTAAGTTGAAACGTTTTATCGG
WTHIN24731_5R	23272 > 23246	AGAAGATAAAAAGGATTGCTGTAATCG
WTHIN_29208_F	29208 > 29228	TGGTACAATGTTCAAGGAGGC
WGAP50_F	30429 > 30449	CTGTTTGCATCATCATGGTGG
1740_F	30466 > 30485	CACCACACTCATCATTGTCC
WTHIN_30489_R	30489 > 30468	ACAAGGACAATGATGAGTGTGG
WGAP50_R	30807 > 30780	GCACTGAAAATCAAATATAGACAGAACC
WTHIN24731_6F	39335 > 39356	TGTAACTCTTGGGTGAGATTGC
WTHIN24731_6R	40585 > 40561	AGTCGAAGGAATACTATTTGTAGCC
WTHIN_40819_R	40819 > 40795	AACAGAGGTATTACAATACATTGGC
WGAP51_F	43727 > 43758	AATCAATTAGTGGATATTATAGATACTGATGG
WGAP51_R	44893 > 44869	CAACTGGTACATTCAAATCTACACG
WTHIN_45092_R	45092 > 45067	CCATCTACAAATAGCACATTAGTTGC
WTHIN_45283_R	45283 > 45261	CTATAGATGCATGTTTACGGTGC
WTHIN_45361_F	45361 > 45385	TCGAATAGTGATAAACCAGTAGGTG
WGAP52_2F	45391 > 45414	GTAGGAGGTGCAAAAATATCAACG
WGAP52_F	45865 > 45886	GGTGGAGGTGACAGTATATGTG
440_LE	47113 > 47083	CGAAGGATAAGTATTATAGTAAATTGACAGC
WGAP52_R	48262 > 48236	TTCCAAGTCATATAGTCATATTCCCAG
WGAP52_2R	48382 > 48361	AAGGTCTGTTTGATCAGTCTCG
WGAP53_F	57820 > 57849	TTTGTAGTAATTCTGTTACCATCTAATGTG
WGAP53_1491R	59015 > 58987	CTATATCTCATGTCACAACTATCATATGC
WTHIN_59186_R	59186 > 59160	CAACAACAATTTGCTATTCTTGTTCAC
WTHIN1491_1F	60579 > 60605	AGTATTACACATAGGAAGAGTAAGCTG
WTHIN1491_1R	60811 > 60788	TCAACAGTTGCAAAATTTATCCCC
WGAP53_1491F	61996 > 62019	TGCTTGTGATTTTATTGTGACAGG
WTHIN440_1R	62001 > 61976	CAAGCACTTCTTGTTAATAGCTATCG
WGAP53_R	62593 > 62570	AGTCAACATCCTATCTCAAGACTG
WTHIN14674_1F	67360 > 67384	AGTAATTGCTAATCAATGCAAGAGG
WTHIN14674_1R	67682 > 67659	TCCAATACAGTTAACTTACAGCCC
WTHIN14674_2F	79317 > 79344	AATATATCTGGTCTTGTTTCACATAAGC
WTHIN14674_2R	79903 > 79879	AAAGAACCAACCTTAACCTAATGAC
WGAP59_F	80643 > 80670	ATATTAAGCAGCTAAAGTCTTTGTAGAG
WGAP59walk_1F	81166 > 81195	TGGTATGTTAAAGATAAGATGGTTATTGAG
WGAP59walk_1R	81770 > 81744	TTTTACGATAGACATCATAACCACATG
WGAP59_R	82364 > 82335	CATATAACACAATAGTAATAAGGAGGACAC
WTHIN440_2F	87528 > 87550	TCCCTCCAGATTTGATTAGATGC

Primer name	Position	Primer sequence (5' > 3')
WTHIN440_2R	88170 > 88145	AGAGACTTACATTTACATCCACATCC
WTHIN_88855_R	88855 > 88834	CCAGCATGATGAGTTAATACGC
WGAP60_F	92316 > 92344	ACTGATGAAATGATAGAGATACTAACTCG
WGAP60_R	93556 > 93528	GCATTCGTTAATCATAGTTTATTACATGC
WTHIN440_3F	111299 > 111325	TCTATGAAAGATAAGCTAAGTGATGGG
13618_RI	111763 > 111789	GTGTAATATAAGTGTCAACTGAAGGAG
15457_LE	112323 > 112301	CAACAACAGATCTAAGCTGAACC
WTHIN15457_1F	112728 > 112756	GTATTGTAATACAGAAGCTCAAGTATCTG
WTHIN15457_1R	112977 > 112954	ACCTAATACAGAACTATCAGCACC
WTHIN440_4F	116709 > 116730	GTGCATCAAGTACATCAGAAGC
WTHIN440_4R	117416 > 117390	ACATCCATATAAGTCTCTTATCACAGC
15457_RI	123395 > 123417	TGCAAAATAGAAGGAGAAGTGGG
WTHIN440_5F	124111 > 124136	CTTGGAAAATCAACTTGTATGATGGG
15288_LE	124137 > 124114	ACCCATCATACAAGTTGATTTTCC
WTHIN440_5R	124906 > 124882	AGCTACCTTTGACATTTATACCTCG
WTHIN15288_1F	127329 > 127351	TGTACTGTGTGATACTTGGAGTG
WTHIN15288_1R	128211 > 128185	TCTACCTAATACAATAACAACAAAGCG
WTHIN15288_2F	129703 > 129725	CTGCAGTTATGATAAGCAAGGTG
WTHIN15288_2R	130292 > 130270	GCATAGCAAACTACAGTCACAGC
WTHIN440_6F	134161 > 134183	GAAAGGAAACAATGACATGGGAG
WTHIN_1F	134315 > 134343	GTGGAATATTTTAATAATGGACAAGATGC
WTHIN_1R	135640 > 135617	AACATGTCCTATATGTAGTTGCCC
WGAP19_F	138617 > 138645	CAAATATTGTATTGATAATTCACTGTGCC
WGAP19walk_1R	139723 > 139702	CATCCATTAGTAACCATGCTGC
WGAP19walk_2R	140593 > 140567	GATTTCAGGTAATATGAAGAATGACGG
WGAP19_R	142129 > 142105	CTGATGACATCAGGTCTTTATTGTC
WTHIN912_1F	178258 > 178282	CTACATTGCACACATACATCATAGG
WTHIN912_1R	178733 > 178708	AGATGATAGATTGAAGACCTTAGCAC
WTHIN_183502_F	183489 > 183511	GATGTTCCTACAGTAACCAAAGC
WTHIN_184516_R	184503 > 184482	CACATGCATGAACTACAACTGG
WTHIN912_2F	194799 > 194821	TGGTAAGTGAACTCTTCAAGTGC
WTHIN912_2R	195553 > 195533	GACAGGAAATAACAAGGCTGC
WGAP18_F	199220 > 199246	TATTGACATTCAATTCGGAAATATGGG
WGAP18walk_1F	199598 > 199619	TGTTTGGTAAGTGTAGGAGAGG
WGAP18walk_1R	200616 > 200590	TCCAATCATATCAAAACACAACATCAG
WGAP17_F	201214 > 201189	CCACCTAAATCTTCATCATTGATACC
WTHIN_204439_F	204426 > 204448	GGTGAACCACTTGTAAACATTGC
WTHIN24663_23R	205295 > 205320	GTTAAAAGTAGGACTGCTGTATTTGG
WTHIN24663_23F	206059 > 206037	TGCACAAGTCTAACAAGTCACTC
WTHIN24663_22R	208735 > 208759	TGTTGATGTAGGATTTTGTTATGGC
WTHIN24663_22F	209611 > 209585	ATACTCGTTAACACTTATTCTAAAGCC
WGAP18_R	217387 > 217409	CTTTGAGCTATTAATGGTACGGC
WGAP17_R	218500 > 218471	ACATTTCAAAGATAACAAATCACAATATCC
WTHIN24663_20R	225283 > 225311	GTTATGTTATATCTATGTGCGGTTTATGG
WTHIN24663_20F	226046 > 226019	ACTAGATTCACACAATACATATCTCTCC
WTHIN24663_19R	230805 > 230829	TGCTCATACTTTTGAAATTCAGTCC
WTHIN24663_19F	231296 > 231269	GTAATACTAGAAGAATTATGCACTGTCG
WGAP37_F	236390 > 236414	TCATGTAGGAAAGTTTTGTGTTGTG
WGAP37_R	237392 > 237361	ACCATAGAACATTTCTTTAGTAGTTATATTCC
18484_RCF	241513 > 241539	GTAGTGATAGGTTTGTAGTGTTAAGTG
WTHIN24663_15R	243146 > 243172	GATTTGTAGTTTTGGTCATACATGAGC
WTHIN_243656_F	243643 > 243665	TGGTATGTGAGTATTGCGATTGG
WTHIN24663_15F	243666 > 243644	ACCAATCGCAATACTCACATACC
WTHIN_244515_R	244503 > 244483	GCAGCAAGCTATCAAGACAGG
18484_RCR	244644 > 244614	GATAGTACCATATTCTCTATCATACTTACTG
WTHIN24663_13R	255834 > 255860	CCCTGGTTTATCTAAATATGGTTTTGC
WTHIN24663_13F	256412 > 256386	AGGAGAAACAATGATTGTATTAATGGC
WGAP38_F	269703 > 269731	ATAGCTATGTTATAAGGTGTAATTGAGTG
WGAP38_R	270006 > 269979	CACATTACCTTTTGCAACTTATAAACAC
WTHIN24663_11R	270649 > 270678	GTTAGTTAGAACTAGTCTGATAACAATTCG

Primer name	Position	Primer sequence (5' > 3')
WTHIN24663_11F	271698 > 271674	GCTGCATCAGTATATCTTTCATCAC
WGAP39_F	274371 > 274393	TCGGATACACTAAGAACAACTGC
WGAP39_R	274762 > 274742	CGCTATCTGGAACTTAGCAGG
WTHIN_276069_F	276055 > 276077	ACACTATGCTCTCTATGTGATGC
WTHIN_276967_R	276953 > 276930	CAGAGTTGCTATATCCCTATTCGC
WGAP40_F	279398 > 279424	CACTGTAAGTTTTGGTATTTAGATGGG
WGAP40_R	279840 > 279811	ATTTGTAGCATATAATACTATCAGTAGCAG
24993_RCF	282977 > 282999	CGATCTATGTCTCAAGGTAGAGC
WGAP41_F	283515 > 283539	TGAAGAGATGCTATCGTTAGTTGAG
WGAP41_R	283925 > 283905	AATATCCCAGCATTATCCCCC
24993_RCR	284382 > 284361	AAACTATGGCAGGAGTGATAGG
WGAP42_F	285508 > 285528	TGCTGATACAGTAGATGCTGC
WGAP42walk_1R	285933 > 285959	TTTTATGTCTTCTGTCTCTTCTATTGC
WGAP42_R	286888 > 286865	CTCCATCTTATCTACTAGTTCCGC
WTHIN24663_9R	295986 > 296011	GAAAGTGTATGCTGATGTATTAAGCC
WTHIN24663_9F	296960 > 296931	CATCATATCTAGTAACTTTAGGTAGCTCTC
WGAP103_R	299058 > 299085	TTATAATTCTATGTGGCTAGTCTTTTGG
WGAP103_F	299445 > 299418	GCACTTAAACACAATTGAAACTTTTGAG
WTHIN24663_8F	300242 > 300218	GAAACACTTCATATACAGTACCCAC
WTHIN24663_7R	303780 > 303802	ACCTCTACTAAGACTGAGAGCAG
WTHIN24663_7F	304262 > 304239	CATATTTCGACCTATTTCTGCCAC
WGAP102_R	313871 > 313894	AGCGATTTGTAAATGTGTGAAACC
WGAP102walk_1R	314264 > 314288	TGCTTGATCAAATGAGATTGATTCG
WGAP102walk_1F	315065 > 315035	GTACTATAGTTGAGATAACGAACATTAAGTC
849_F	315245 > 315209	ACAATTGATACCTAAGTAGCTACAGTC
WTHIN_315739_R	315726 > 315704	AAACTACCTACTGAACTACCAGC
WTHIN23036_1R	318174 > 318200	AGGCATTATTATTATGTTCTTGTTGGG
WTHIN23036_1F	319336 > 319309	TTGTAGCATGTATTATTAGATCATCAGC
849_R	339626 > 339654	CTTGAACATACATACCACATATACCTACC
WGAP101_F	340493 > 340467	GAAGTTGTTATTGATGAAGTCATAGGG
WTHIN24663_5R	346090 > 346113	TCAGAAATAAGAGGTCATCGTAGG
WTHIN24663_5F	346592 > 346570	ACTGCCTTTCTCATAAGACTAGC
WTHIN24663_4R	348908 > 348931	AGACAACAGTATCTTGAGCATACG
WGAP100_R	349303 > 349332	GTGATTGTTAAGGATATTATTCTATGTTGC
WGAP100_F	349809 > 349784	GCAAGCTTGATACTTGTTAATCTGTC
WGAP99_R	350692 > 350716	TTGTAGTACAGATGGAAGGTAAGAC
WGAP99_F	351186 > 351160	CGATAAAAAGTTGAAACAACGTATCCC
WGAP98_R	351426 > 351450	TGAGGAAAGAGTTAGTATGCTTAGG
WGAP98_F	351691 > 351665	AGTACAATACATCGCTATAATAGGGTC
WGAP97_R	352556 > 352582	TTGTGTGGTCTTGTATTAATAGTTACG
WGAP97_F	352862 > 352840	AGATTCTGTCCATCATCATGAGC
WTHIN1639_1R	354298 > 354323	TGCAATAAAATATAAGGCATATGGGG
WGAP96_R	355056 > 355082	TGTGAGATAATGACTTAACAATATGGC
WGAP96_F	356303 > 356279	CTAAGCAGTATATTGGATCTTGCAG
WTHIN24663_3R	356595 > 356618	CAATATTCCTTGTGGTCCTAGAGC
WTHIN24663_3F	357305 > 357282	ACTCTTTACCCAAAGTAGTAACGC
WGAP95_R	360602 > 360623	AAAGGTACAGTAACTGGAGAGC
WGAP95_F	361685 > 361662	GTTATTACTCAGTACCGCATCTGG
WGAP94_R	362643 > 362668	CCATCATGACCATATAAGTACACTCC
WGAP94_F	364025 > 364001	ACAAGCTTTACTGTTCCATTTTCAG
WGAP16_F	371152 > 371178	ACATGACAAGATCTACAAAATCAAGAC
WGAP16_R	372094 > 372066	AGGAACAATTTGCAAATTCTAAAATTGAG
WTHIN24663_1R	375188 > 375215	CAACATGTGTAATTGTTACAGATTGGAG
WTHIN24663_1F	376135 > 376113	TCTTGATAACAGATTGCCTCCTC
WGAP15_F	384394 > 384421	TGAATTGAAGACTTGATATGTATTCCTG
912_RI	385013 > 385039	CTTTTGTGTAAAGGAGTATGTACTAGC
WGAP129walk_1R	385855 > 385835	TCATCAGCACCAAAAACAACG
1174_RI	386308 > 386289	GAAAACAGCACAAGGCAACG
WTHIN1174_1R	387585 > 387611	CTTGTAATATGGTCGTTGTAAAATCGC
WTHIN1674_16F	388823 > 388796	CATATGTTAAGCTATATCTATGCACCAG

Primer name	Position	Primer sequence (5' > 3')
1174_LE	389080 > 389107	CTGGAGGTTTCTTTTATTGTACTATCTG
WTHIN1174_1F	389107 > 389080	CAGATAGTACAATAAAAGAAACCTCCAG
776_RI	389955 > 389926	GTAACCTAATAAATAAGTACTCTCTCAACG
WTHIN1674_15R	391106 > 391130	GGTGTGTTCATTGTTTTGAGATAGC
776_LE	392597 > 392623	AGTAACTTGATATTTTGCAGTGTAGTC
WTHIN1674_15F	392663 > 392634	ACTACTCATTTTAGTACAACTAAGTAGGTC
2808_RI	392802 > 392774	GTATGTTATGCAAACCATAAACTATTGAG
WTHIN2808_3R	393364 > 393388	AGATTGAATGAAAACAGTGTTTGGG
WTHIN2808_3F	393891 > 393872	ACCACCAACAACAGATACCC
WTHIN1674_14F	393933 > 393910	GAAGAGAGCATTTAACAACATCCA
WTHIN1674_13R	397364 > 397390	TGGTAGTTAGATATCATGCTGTTTAGG
WTHIN2808_2R	397365 > 397392	GGTAGTTAGATATCATGCTGTTTAGGAG
WTHIN2808_2F	398070 > 398045	CATGTTCTCCAACTATCTTAGACACC
WTHIN1674_13F	398184 > 398159	TTACTTCAATACAAAGCTGATTTGGC
WTHIN1674_12R	399636 > 399658	AGAGTCAGACAACTTAAGCATGG
WTHIN2808_1R	399751 > 399775	CGTGTTCTTAATGTACCAAGTTTCC
WTHIN1674_12F	400783 > 400763	CAAGTTGTCATAGCCTTGAGC
WTHIN2808_1F	400786 > 400764	CAACAAGTTGTCATAGCCTTGAG
2808_LE	403289 > 403311	CCGTGTTCTACGTTTAGTGTTCC
WTHIN1674_11R	403995 > 404018	TCAAAAGTATCACATGTTCACACG
WTHIN1674_11F	404987 > 404964	GCATATTAGCTGATAAAGGAACCG
WGAP14_F	405892 > 405868	AGCCTAATTCTAGATCATCATCACC
WGAP15_R	413104 > 413125	TGTTGGTCGATCAAAGACTCAG
WGAP14_R	413667 > 413640	TTTGATGTCTAGTTCATACATTTACCAG
WTHIN1674_10F	414450 > 414424	ACCTTATAAAATCAGCGAACTATAACG
WGAP13_F	417919 > 417945	TGCAGTTAACGATATTAGAATTGTTGG
WGAP13_R	418527 > 418501	TTACAAGACCTACTCTATTACTAACCC
WGAP12_F	425060 > 425086	AGGTGTTCAATAAGGTTGTAAATACTG
WGAP12_R	425275 > 425250	GTACACTTTAGGACATATAACACTGC
WTHIN1674_9R	427161 > 427187	GATAGTATCTGTGTTGTGAATATGCTG
WTHIN1674_9F	427951 > 427925	GTACCTACTTAAACATAACACTCATGC
WTHIN1674_8R	436561 > 436586	CTTCTGTAGCTCTTCTATAGTTCTCC
WTHIN1674_8F	437056 > 437035	AAGCAACAATGACGAAGTTACC
WTHIN1674_7R	462390 > 462413	ACAAGTATTTATTGCACACAGGTG
WTHIN1674_7F	463320 > 463302	CAAAGCGAGTAGGTGGAGG
WTHIN1674_6R	464416 > 464441	ACATTGGTATTGTCATACTTACTCCG
WTHIN1674_6F	465134 > 465112	AAGAACTGGAGTAGATGTAAGCG
WTHIN1674_5R	470438 > 470465	CTATGATACAACTTAGACATTTCTGGAC
WTHIN1674_5F	471498 > 471478	ACATCCATCTGCTGAACTACG
WGAP11_F	472684 > 472710	GCTATATCAGCTGATAAACTTGTTGAC
WGAP11_1640F	474112 > 474090	ACATATGCAGCTATGGATGTAGC
WGAP11_1640R	475854 > 475875	TGAGTGTTGTTGTAGTGCAGAG
WGAP11_R	477196 > 477171	CAATAGCATTAGCTTTCTAAGTGGAC
WGAP10_F	480631 > 480658	GATGTTAATTTCTGGGTTCTTATCTCTC
1458_LE	481320 > 481291	GCATATTACTTCCATAAAATCTTCACACTC
WGAP114_F	483116 > 483138	AGGAAAGTGCTTGTATTGTAGGG
WGAP114_R	483560 > 483533	CCGTATAGACTTAGTCATAGATGAAACC
WTHIN1674_4R	483663 > 483686	TCTTCATCTACAGGTTCAGTATGG
WTHIN1674_4F	484973 > 484948	TCTATAAATAGCTCAGTACTGGAAGG
4518_RI	488477 > 488500	TGGATTAAGAAGACTAGCATCAGC
3205_RI	488808 > 488782	GTAGTGCTACTATAAAACCATTACCTG
WTHIN1674_3R	489262 > 489288	GTGATGCTACAAAATCATATACAGTGC
WTHIN1674_3F	490724 > 490704	TCGTCTTTCTAAGGAAGGCTC
WTHIN1674_2R	491999 > 492022	TGCACCTAAGATGTATAAAGTGCC
WTHIN1674_2F	493588 > 493562	CAAGATAAAGCTATACCTATTGAAGGC
3205_LE	493914 > 493939	GGATAGTTCATTAATTGATGGTCTGC
WTHIN21267_2R	494746 > 494770	CTGAGCTTGAAGATATTGTTTACCG
WGAP10_2178F	494793 > 494768	CATATAGCATTTCACTAATGCATCGG
WTHIN21267_2F	495472 > 495449	GCACTGCTAGTACTGATCTTAACC
WTHIN24706_33R	495771 > 495794	GGTATTGTTACGTACTTTTCAGGG

Primer name	Position	Primer sequence (5' > 3')
WTHIN24706_33F	496233 > 496207	ACTATACATAGCAATACTTACTGGCTG
WTHIN21267_1R	497454 > 497480	TCATTTAGTATAATCAGTGCATGATGG
WTHIN21267_1F	498293 > 498270	CATACAGTTTAACGCTAACACTGC
WTHIN24706_32R	499133 > 499157	CGTAATTCTATAGAAGAAAGCAGGC
WGAP10_2178R	500006 > 500033	AAATTTATGTTCAAGATTTGTTGTGCTC
WGAP10_R	500889 > 500860	CAACTATCATAGATAAAATAACAGCTTTGG
WTHIN24706_31R	508762 > 508784	TCTAGTATTGGCATAGTGGTGTG
WTHIN24706_31F	509280 > 509259	CAAAGCAGTCACAAGATACACC
WTHIN24706_30R	515037 > 515059	GCTATTGGATGATTTGGAATGCG
WTHIN24706_29R	515301 > 515324	AATGCATAAACAGTTTCAGTAGGG
WGAP93_R	516221 > 516250	GAATATATGACCTCAGCTAATACTAATGTG
WGAP93_F	516699 > 516679	TCCTATTCCACCTGTCAATCC
WTHIN24706_28F	517019 > 516996	TCTATCAACAGAGGAACATCAAGC
WTHIN24706_27R	517288 > 517313	TTGAGTAGTGAGGTAATTTCTAGAGG
WTHIN24706_27F	518543 > 518517	CTCTAAGTAAAGTACTGAACTTTCAGC
WTHIN24706_26R	520785 > 520810	GTATGTATTCTTTACGTGAAGATGGG
WGAP29_R	522165 > 522191	ATTTTTATGCTTTACTACTGGAGATGC
WTHIN24706_26F	522276 > 522256	AGTTGAAGAGTGTATGGGACG
WTHIN24706_25R	523580 > 523605	TCGTAAAGAATCATTAGTCAATTGGG
WGAP29_F	523671 > 52365	CACCAGAACGCCATCCT
WGAP30_R	524141 > 524158	TCCAGCACATGATCAGCG
WGAP30_F	524641 > 524620	CCACAAAATGCTGCAAAATACC
WTHIN24706_24R	526080 > 526104	TCAATGAAATTAACACATCAGCTGC
WTHIN24706_24F	527864 > 527841	GGAGTAAAACATGCAACTTCTTGC
WGAP31_R	527897 > 527921	CATTTTTATCTCCTGAACCATACCG
WGAP31_F	528589 > 528560	ATTACCTCTTAAAAATTTACGAAACATCAC
WTHIN24706_23R	529007 > 529036	GTATCAAATTGTGTAATAGTTAAGCTTTCG
WTHIN24706_23F	529990 > 529966	GCAACTCTTAATTGTGTAAATCCGC
WGAP32_R	532303 > 532328	GTGAGATTTTAGCTAAGCATGATGAG
WGAP32_F	532681 > 532654	TAGTGATATCATTCAACACAATACACAG
WTHIN24706_22R	533631 > 533659	GATTATCTCAAATCTAGCTTCTCTATGTG
WTHIN24706_22F	535193 > 535164	TCAGATTACCATAAGTAATAATCTACCCAC
WTHIN24706_21R	545074 > 545099	TGTAAAAGTGAGAACAGAGTCTAACG
WTHIN24706_21F	545885 > 545869	TGGGGTTATGTGCTGGC
WGAP33_R	546541 > 546560	CTGCAAGAGATTGCGAAACC
WGAP33walk_2F	547678 > 547653	ACACACAAACATATGTACAAAAGAGG
WTHIN_547962_R	547946 > 547922	CAAATAACAAAGACGATAGAGCACC
WTHIN24706_20R	548257 > 548279	GACAGCATTGTTTGTTGTGATCG
WGAP33walk_1F	548446 > 548470	ACACAGTGGTAAACTATTGTCTAGC
WGAP33_F	549283 > 549264	AAAAACCAAGACCAACCTGC
WTHIN24706_20F	549599 > 549578	ACCAAGAAGAACACAAGACCAG
WGAP34_R	552572 > 552596	ATAAAAGTAGATTGTTGTGGGTAGC
WGAP34_F	552975 > 552947	CTACATTACCTTAATAACATTCACCACTC
WTHIN24706_19R	554517 > 554538	GGAAGGTTATGTGTAGCGATGG
WTHIN24706_18R	555548 > 555571	AGCGTGTAATTTTGATGCTTTGTC
WTHIN24706_19F	555572 > 555549	TGACAAAGCATCAAAATTACACGC
WTHIN24706_18F	556494 > 556475	TGCATATAAGACGGCATGGG
WTHIN24706_17R	563249 > 563273	CTTGAAACAAATTGTCGTCTTCCTG
WTHIN24706_17F	564204 > 564184	GAATCCACGGAGTTTGAAAGC
WTHIN24706_16R	566107 > 566130	AGGCGTGATACTATATTTTGAGGG
WTHIN24706_16F	567494 > 567475	AGCCTAACAGACAATCACGC
WTHIN22116_1R	567809 > 567836	GTATTCTTCCATTGATTATAACACGACC
WTHIN22116_1F	568135 > 568109	CATCAAATCCAAAAGTAGACTATACCC
WTHIN24706_15R	570298 > 570323	GTTTACGTTTGTAATGCTCTTCAAGG
WTHIN_570833_F	570821 > 570845	AGAGTACACTATCATTCTGTTGGTG
WTHIN24706_15F	570846 > 570822	ACACCAACAGAATGATAGTGTACTC
WTHIN_572096_R	572084 > 572058	TGAATACATGAACCAAATTGAAGAACC
WTHIN24706_14R	572749 > 572771	GGTAATTTGCTGTATGTCTTCGC
WTHIN24706_14F	573401 > 573378	CACTTATTAATGTTCCTCTGCCTG
WGAP35_R	574850 > 574875	TAAAGTGTATCGCAATATTAGTTGCC

Primer name	Position	Primer sequence (5' > 3')
WGAP35_13F	575449 > 575427	TGTATAACATCACCACTTGCTCC
WGAP35_13R	577433 > 577460	TGTGTACTAATTGGTAATGCATTATTGG
16700_F	578018 > 577991	ACAGCTGTATTTACCATTTAAACATTCC
WTHIN24706_13R	580367 > 580391	TGTTTAAGGAATGACAAATCTCACC
WTHIN24706_13F	581133 > 581114	ACATATAGGCTGGTACAGCG
WTHIN24706_12R	582829 > 582851	CAGAACACGTGATACAAATTGCC
WTHIN24706_12F	583305 > 583281	GAACATCACTACCAATCTCTAAAGC
WTHIN24706_11R	585166 > 585186	CTGTCCTCCAGTTTCCATAGC
WTHIN24706_11F	585476 > 585458	AACACTCCACAACAGACCC
WTHIN24706_10R	586251 > 586273	TCAGTACGAAGTATTCTTGAGGC
WTHIN24706_10F	586830 > 586802	GTCATAAGAAGATATGGATTATCAGTAGC
WTHIN24706_9R	593237 > 593262	TGAACAAATGTCTTTAGTTGATGCTC
WTHIN24706_9F	594158 > 594133	GCATTTTGGTAAGGACATCTAATTCG
WGAP36_R	595679 > 595706	TTAGTGTGAATGTGGTTATATAACAGTG
WGAP36_2F	596990 > 596968	ACATGGTCTCATAGAAGTTAGGC
WGAP36_F	597419 > 597399	AGAAACTTGCCCTATTCCAGC
WTHIN24706_8F	597747 > 597723	CACGATATGGAATTATGAGATGCAG
WTHIN24706_7R	601012 > 601041	AGCAAATTTCCTATATCTAAGAATAACGTG
WTHIN24706_7F	602281 > 602254	TTAGTAATTTATCATAGAGCTAACACGC
WTHIN24706_6R	606101 > 606124	GTGGAAGTTTTTGTAGTGAGTACC
WTHIN24706_6F	607627 > 607603	AGTGTAGCTTCCATAGATAATCCAG
WTHIN24706_5R	608296 > 608316	AAGTACTGTGTGCAAGGTCTG
WTHIN24706_4R	608726 > 608754	GTTCATCATTATGGAAATAGGAAAAGCTG
WGAP92_R	608978 > 608997	CTGCACAATTTGTTCGGGTG
WTHIN24706_3R	609572 > 609594	TGGGTGTAGGGTTTAATATTGGC
WGAP92_F	609881 > 609859	GCATCAAAAGTAACTCTGCATGG
WTHIN24706_3F	610528 > 610502	GGTACAGATAGAAACCTAAACTGTAGG
WTHIN24706_2R	615392 > 615410	AGATCATGCAGGATGGCAC
WTHIN24706_2F	615900 > 615878	ACCGTTGTTATGTAGATCATCCC
WTHIN24706_1R	617979 > 618001	TCTTCGTGAGGTTCATAATCCTC
WTHIN24706_1F	618250 > 618228	CCCAATCTAAACTTGCATAACCC
WGAP112_R	618458 > 618484	AGACTATATCACCATATCGTAAAGACG
1674_RI	619846 > 619820	CATTATGTCCATCATTAGATTCAGCAG
WTHIN1674_1R	621066 > 621091	ACTTGCTATTCTTTATTCTAGGGGAG
WTHIN1674_1R	622066 > 622086	TGACTCACATGTTGCAGATGG
WTHIN1674_1F	622811 > 622785	GCCTTACACAACTCTTATCATCATCTC
WTHIN1674_1F	622887 > 622865	TTCGGATCATTATCACCTGTTGG
1674_LE	623274 > 623301	AATGGTAAAGATATTTCTAGTGTTGGAC
1519_RI	624641 > 624613	ACTTAGATCCAGAGTTATTATATGACAGC
1519_LE	626131 > 626156	CAGTTTTTCATATGTTGGAAATGTGG
1309_RI	628639 > 628663	GATGTGAACATAAACCATTTTCCCC
2267_LE	629051 > 629028	TGCTGCTGATATAAATACTCCAGG
2267-2_LE	629326 > 629306	ACGATTATGTTGCACAGAGGG
WTHIN2267_1F	629465 > 629486	GCATAACAGCCAATAAGATCGC
WTHIN2267_1F	629596 > 629625	GCAACTATAAAAAAGTGATACACTATACGC
WTHIN2267_1R	630448 > 630420	AATGCTAGAGCATATAATAAATGTACCAC
WTHIN2267_1R	630672 > 630652	AGACAACGTATTGTCTGACCC
WTHIN2267_2F	632078 > 632101	GTAGTTTAAGATAAGCAGCGATCG
WTHIN2267_2F	632462 > 632493	GTATATCTGGAAATTAAGATAGTGGTTAAGAG
WTHIN2267_2R	632919 > 632891	TGCTAATTTCAGCTATTCAAAATACACTC
WTHIN2267_2R	633283 > 633257	CTACTACTACTGTATCATCAGCTTAGG
2267_RI	634044 > 634069	AGTTATTATCGTTGTCTATTGGAGGG
WGAP130walk_1R	634683 > 634704	TTGCACTAAATACAACAGCAGC
WGAP112_2F	636855 > 636828	ACAAAACCAATAACATACTGATAATGCC
WGAP112_1F	637533 > 637507	ACTAGATTACATAACTAGCCACTACAC
1309_LE	641215 > 641235	TCCAGATGACAATGCGTATGC
WTHIN24688_3R	646313 > 646334	CATGTGATACACCATTAGCTGC
WTHIN24688_3F	647161 > 647135	CCCTAAGTATAACAAATGTCTTACACG
WGAP91_R	649405 > 649431	AGCATTATATGTACATAAGCATCAACC
WGAP91_F	650879 > 650853	ACCTATAATAGCATACAGTAGGTTCAC

Primer name	Position	Primer sequence (5' > 3')
WTHIN24688_2R	652228 > 652252	ATGTTTCTAAGTAACGACAGATTGC
WTHIN24688_2F	652853 > 652832	GTGTTTTCGACGAAGAAACAGG
WTHIN_657239_F	657227 > 657249	CCAGAATGTGTTGCTACAATACC
WTHIN_657714_R	657702 > 657679	ACAGACTTATTAGCAAGTAGTGCC
WTHIN_660645_F	660633 > 660660	TGTTCTTAATCAATAATCATACCACTGG
WTHIN_661418_R	661406 > 661384	CCTCCACTTTGTAGTTTCTCACG
WTHIN1360_1R	662830 > 662854	CAATTCCTATACCATGTTTTCCACC
WTHIN1360_1F_C	663088 > 663063	AATGAACCACCATATTCAATAGATGC
WTHIN1360_2R_B	663751 > 663773	TGTAGCTAGCTGTACAATGAGAG
WTHIN1360_2F_B	664421 > 664397	ACCCAAACATTAATACTGAAGAACC
WTHIN1360_1R_A	665086 > 665112	AGTTTTGATGCATCATCTAATTCTGTC
WTHIN1360_1F	665357 > 665328	AAAGATACATTATGATTAGTAATCCTCTGC
WGAP90_R	666007 > 666032	GACTGTAATGAAATGTCATGACATGG
WGAP90_F	667254 > 667228	CATACACAACCGTAATTTACTACTCTC
WTHIN24688_1R	667872 > 667894	GTCTGTTAGTGAGAATTCTTGGC
WTHIN24688_1F	668421 > 668401	CCAATTAGCATCTTCACCTGC
WGAP89_R	669482 > 669502	TGCCTTCTGATCATGTTCCAG
WGAP89walk_1R	669799 > 669822	TGAAGATTCAATGACACAATCTGC
WGAP89walk_2R	670220 > 670241	AGCATTCTGGTGTAAAATGCTG
WGAP89walk_2F	671093 > 671073	ACTTGCAAACTCAACAACACC
568_LE	671677 > 671651	CTATGCAGCATAGTATTATCAGTTTCC
WGAP89_F	673249 > 673229	CAAGGTAAGAACAACAACGGG
WGAP88_R	679057 > 679082	GTTAAATCAGTTACAGTAGTACTGCC
WGAP88walk_1R	679774 > 679800	TGAGTAAGGGTAATTATTAATCGAGGG
WGAP88walk_1F	680482 > 680453	ACCTGCTTATAATTTATTATTGTCTACCTG
WGAP88_F	680960 > 680934	AGGCATTGAATCAAATATATGGATACG
WTHIN_681369_F	681357 > 681336	CTAGAACTGGCAGAATGTTTGC
WTHIN19567_1R	681478 > 681506	GTTCTAAAATTATAGCAGCAGTACTATCC
WTHIN19567_1F	681958 > 681935	GATGGTATATCTAGTTGGTGGAGC
WTHIN2007_1R	682328 > 682353	GACTGTAATGCTTAGCAATTGTAGAG
WTHIN2007_1F	682591 > 682565	CCCTTATACCTTATATTTGTACATGCC
WTHIN_684129_R	684135 > 684117	GTCGTTGGAAACCAAGTGC
WTHIN_685216_F	685204 > 685226	AAAAGCTAAGGAAATCATGAGGC
WGAP87_R	686334 > 686357	GAGAGCAGCATATGTATGTTATGC
6654_F	687162 > 687134	GTACAACATTAGTAAGTTCTGAAAAAGTG
WTHIN568_1F	690695 > 690717	GTTTAAAGTTGCTAATGCTTGCG
WTHIN568_1R	691164 > 691140	AAGGAAATATAGGGTTAATGCAAGC
WTHIN_699689_F	699677 > 699702	GGAAGTTTACTTAGTCCTGAAGTAGC
WTHIN_700846_R	700834 > 700814	AACATGCTTCATCTGTCCTGG
WTHIN_706368_F	706356 > 706379	TGTGCTTCTTATCATCTTACGACG
WTHIN568_2F	707467 > 707488	TCTAGTGGGAAGTACACTGGAG
WTHIN568_2R	708059 > 708039	GTCAGTTGAAGCAATAGCAGC
25104_RCF	709379 > 709401	AGTATTGATCATATCTGGCGGAC
699_RCR	710214 > 710188	GTATGCTGGTAAATCTTCTTATTCCAC
699_RI	710687 > 710714	AATTACTACTGCTAAATAAAACGTAGCC
WGAP134walk_1R	711359 > 711384	TCTATAGCTGTATGCCTCATATATGC
WGAP134walk_1F	711965 > 711943	CAAACATGGCATGATAAACTTGC
10652_F	712830 > 712803	TGATAGTCAATTGTATAGTTAATGTCGC
WTHIN878_2R	713745 > 713769	GTAGAGAGAACTAGAAAAGGAGGTG
WTHIN878_1R	714794 > 714824	AGTATCTGATATGTGTTTTAGTATGGTAGAG
WTHIN878_2F	714829 > 714799	TGTATCTCTACCATACTAAAACACATATCAG
WTHIN878_1F	716011 > 715984	TGAACAATACAAAATCTTATACGACAGC
WTHIN_727336_F	727324 > 727349	CTAAAGTAACAGATAAAGAGCAGCTC
WTHIN2698_5R	727621 > 727642	TCTGAAATGGAGTTACGTGCTG
WTHIN2698_5F	729140 > 729117	CTATGAGTACAGAATGGATCAGGG
WTHIN2698_4R	730213 > 730238	AACGATTTAAGGTCTTGGTTTAAACG
WTHIN2698_4F	731301 > 731278	ACTCATTTCTGAAGCAGGATAACC
WTHIN2698_3R	731560 > 731587	ACTGGTATTAATGATAGTCCATTACGTC
WTHIN2698_6F	731717 > 731694	ACTATACTGCAAAAATGAGTTCGC

Primer name	Position	Primer sequence (5' > 3')
WTHIN2698_3F	732756 > 732730	CAATGAATCGAATTTGTTGTTGTTAGC
10652_R	740548 > 740575	CAAATAAGCTATGTTTTCTTATAGCTCC
WGAP85_F	740732 > 740712	AACACCAAGCGAAGAAATTGC
WTHIN2698_2F	741862 > 741838	CACACAGTCTTTATATGTTGAAGGC
WGAP84_R	743171 > 743194	CATGTATACCTAATGCATGATGCC
WGAP84_F	743490 > 743463	ACATAGTGATACAGAAGGTATAACTGAG
878_LE	744134 > 744162	CATTACTAAAGTACCTCTCTTAAAATCGC
687_LE	744928 > 744903	CAGAATCCTTAGAATCTTATGATGCC
WTHIN282_3R	745378 > 745397	GCACTCTTGTAACCAACAGC
WTHIN282_3F	745944 > 745919	CATACTACTTCAACTTCAGGTAGCTC
WTHIN2698_1R	746426 > 746448	TCTACAGAATCCCTGTTTATGGC
WTHIN2698_1F	747184 > 747162	TCAAGTGTACCAACTTTTCCTCC
WTHIN687_1F	748247 > 748275	TTAGAAGATAATTCTGGAACTAAAGAAGC
WTHIN687_1R	749165 > 749141	TGGAGCTAGATAATCTTGGTAATGG
WTHIN687_2F	750693 > 750717	CTCAAAGGAATACAAGACACTATGC
WTHIN687_2R	751338 > 751313	CAGGTAACGTTTTACACATATTGGAG
687_RI	751583 > 751604	AGATCATCACGACTCATACCAG
3395_RI	752565 > 752539	ATCCATAGATCAACTTGTTTATTAGCG
3395_LE	756698 > 756723	TGAAGATGATATAGATACACCAGCAG
WTHIN8190_3R	757306 > 757324	CTGAACGTGCTTGGTTGTC
WGAP108_F	757331 > 757311	TACAATTGACAACCAAGCACG
WTHIN8190_3F	758853 > 758821	CACAACTAAAAGATAACATACTGTTTAAATCAC
WTHIN8190_2R	763133 > 763156	ATCCAAAGCCATTAATTCCAAGTG
WTHIN8190_2F	764033 > 764004	GCATATGAAATACTGAATACGAATGATAGG
WTHIN8190_1R	764505 > 764525	AACCCAAGTGGTGATAGTCTG
WGAP111_R	765630 > 765597	CATTCTTTATAATTAGGTAAAGGATATGTATCAC
2546_LE	766075 > 766049	TTGTTCTATTTTAGATAAACCACCACC
WTHIN_773698_F	773686 > 773707	AGAGAGTACCAGAAACTGAAGC
2546_RI	773844 > 773863	ACAACTGACAGTGCTGTACG
WGAP135walk_1F	774156 > 774181	GAAGCATTAATAGCTGAAGATGAAGC
761_LE	775710 > 775685	CTTTGATCTGGTAAGGATAGTAATGC
761_RI	776335 > 776361	CACAATAATAGACATGTTTGAAACAGC
WGAP135walk_1R	776980 > 776957	TGTTTTTATCCTGTAGTTGTGTTGC
WGAP111_F	777769 > 777744	GCAGTTTATAGCATCTATTAGGTCAC
WGAP110_R	781450 > 781476	CTACACACATAACAACAATAGTTAGCC
WGAP110walk_1R	782077 > 782106	CCAAATTACTACATCAAAATACCTAACCAC
WGAP110walk_1F	782818 > 782789	TGTAGCTCTATATATAATGTTCTGATGTGG
WGAP110_F	783197 > 783178	AGTTGGAGTCAAGATGAGCG
WGAP109_R	791387 > 791416	AAGTCATAATGCTATGATCTTAAGTATGTG
1932_LE	792365 > 792344	TGCTGTTTGAGGTAAAGTCTGG
1932_2RI	795195 > 795219	TGTAATTTCTCTATACGACCGATGC
1932_RI	795400 > 795427	CATAGAAATCTGATACCATCTTACTTCG
WGAP109_F	796803 > 796778	AGCAATTGAGAGAATTTTTAGTCCAG
WGAP108_R	799303 > 799331	TCAATAGATTCAATACTCACTGCTATAGG
WTHIN8190_1F	800281 > 800251	TCTTTATAATTAGGCAAAGGATATGTATCAC
WTHIN8190_4F	800611 > 800587	CAGTAGTAATCAAGCAGTAACTGTG
WGAP83_R	804671 > 804691	ACCTACACGAAATACTTCGCC
WGAP83_F	804977 > 804950	GCTTATTACATTAACAGAAGCTAAGCAG
WTHIN12708_7F	805423 > 805401	TGATAGGATGGTTTGCTACTAGG
WTHIN_808166_F	808154 > 808173	TCCTGTGGTAACTTAGCACG
WTHIN_808749_R	808737 > 808717	TGCTGCTGTATTAACTCCACC
WGAP82_R	809362 > 809392	GTAGATTTCATTATACAAATTACCTTTGCAC
WGAP82_F	809740 > 809719	CTGTGCAAGTTACTTCATGTCC
WTHIN_811768_F	811756 > 811781	GAGTAACTTTGTGTGTAGTATCTTGC
WTHIN_813027_R	813015 > 812992	TTTCTAAGAAGAATGTGAAGCACG
WGAP81_R	813322 > 813343	TGCTTTCAAAAGATTCACCACC
WGAP81_F	813899 > 813877	GCTAATGAAATTGCAGTTGATGC
WGAP80_R	819878 > 819903	AAGATAAACATGACAAATACTTGCCC
WTHIN_820343_F	820329 > 820351	TCTGTAACTAAGTCAGCTGTTGC
WGAP80_F	821189 > 821164	AAAGGATTAACACTAAATAAGGTGGG

Primer name	Position	Primer sequence (5' > 3')
WTHIN282_4R	829220 > 829240	GTTTGTAAAGGGTGCTACTGC
WGAP79_R	829676 > 829708	CAGTGCTTTATTGTAATATATACAACATTTCTC
WGAP79_F	830701 > 830677	AATTTGGATGCTAATAACGTAGCAG
WGAP78_R	837879 > 837904	AATTTGATAATGTGAGATGGCAAGAC
WGAP78_F	838322 > 838297	CAAATCCAAGTTCAGAAAGTAGTACC
WTHIN12708_6R	844241 > 844264	CTGGTATATCATGTTTTGGTGGTG
WGAP77_R	845088 > 845116	TAACTTACGAATTAATTTCAAGAACCCAG
16700_R	845673 > 845651	ATGGTGTAACACAGCAATAGGAG
WTHIN12708_6F	846066 > 846046	TGGCAAATGTAGTACCAGGTG
WTHIN12708_5R	847499 > 847520	TGCATCTAATGTAGGAGCATCC
WTHIN12708_5F	848643 > 848618	ACTTGAGTAAGGGTATTTGTTAAAGC
WTHIN12708_4R	849226 > 849251	GCATAATATGGTATATCACGTTCAGG
WTHIN12708_4F	850269 > 850248	CAGCGTCTTACGGATATCTAGG
WTHIN12708_3R	854446 > 854470	AGAATCTGTGCCATCTATATACTGC
WTHIN12708_3F	854970 > 854948	GATGAATTTCATCCTGCATCTCC
WGAP28_F	867655 > 867679	ACCACAATATCTCCAGATGATACTC
WGAP28_R	867882 > 867860	TGGAATTGATGCTGATACTGGAG
WGAP27_F	868593 > 868617	ACCAATCCACAAAATACTAACTCAC
WGAP27_R	869125 > 869096	TGATTAGTCATAATTTATGGTAGAACTTGG
WTHIN12708_2F	869785 > 869760	CCAGCATTGTAATTATCTTGAACTGC
WTHIN282_2R	870812 > 870838	AAGTATAATACACACAATAACTCTGCC
WGAP26_F	870963 > 870994	ATTCATTAAATGGATAAAGTACTTTATCAACC
WGAP26_R	871934 > 871905	GACTATTTGCATTTAGAATGATTGTATTGG
WTHIN3611_1F	875278 > 875304	CAATTAAACAAGCAGCAAATTATCCAC
WTHIN3611_1R	875739 > 875708	GTATCAAAGTGATAAAAACTTTAGTATGGTAG
WGAP25_F	876759 > 876785	AAAGCGTAACATTTAATGTACCTAGAC
4015_RI	877831 > 877807	TTTGCTAACACATTTTACCAGTAGC
WTHIN656_2R	881655 > 881674	TCCACATCTGCACTATCACC
WTHIN656_2F	883235 > 883212	AGTATTAGAACCTGAGTTACAGGC
WTHIN4015_1R	883583 > 883609	TCCTAATAACACATTTTCATCACTAGC
4015_LE	884836 > 884856	GCATACTTGATTGTGCAGCAG
WTHIN4015_1F	884856 > 884836	CTGCTGCACAATCAAGTATGC
WGAP25_R	885290 > 885270	TGTTGCTAAAGGCTATGCTGG
WTHIN12708_1R	887946 > 887968	ACTCTATCAACTGCCAAGATAGC
WTHIN12708_1F	889495 > 889472	GGATGATGGATCAACTAAGACTCG
WGAP24_F	902868 > 902891	AACTCTAGATCTCCTGAAAGTAGC
WGAP24walk_1F	903819 > 903846	TTTATCGGTATTACAACTACTCATTCAC
WTHIN656_1F	903849 > 903824	TTTGTGAATGAGTAGTTGTAATACCG
WGAP24walk_1R	905594 > 905570	AGGAATTTAGCACATATGATTGTGC
WGAP24_R	906184 > 906155	TTTGTAGGTATACATTTGAGTTGATATTGG
WTHIN_907406_F	907392 > 907418	AGAGAGTCAGTATACTACATAACAACC
WTHIN_908473_R	908459 > 908433	CAAAGTATGATGGGTTATATGTCATCG
WGAP23_F	918144 > 918173	TCCCAGATACATTTAATAATGATGATTCAC
WGAP23_R	919004 > 918978	TTACAAGGTAGAGTTTTGTTAATACGG
WTHIN13950_1R	921036 > 921058	TGCTCAAAACTCAGCAATTACAG
24983_RCF	922374 > 922344	TTTCAATATAATAAGTGTATTAGGTTGAGTC
WTHIN13950_1F	922711 > 922684	AACTGATGTTACTATTTTAGGATAGCAC
WTHIN282_1R	923405 > 923430	TGTGCTATACTAGTTCAATAGCATCC
WTHIN282_1F	924122 > 924101	AGTCTTCCAACAACACTTACCG
WGAP22_2R	925773 > 925797	AATACCTACAATCAATCTAGACCCG
WGAP22_F	925921 > 925948	AATACTGATTTACAGCAGATTATAGCAG
WGAP22walk_1F	926342 > 926369	AAACCACTATTAATAGTAAACCACTAGC
WGAP22walk_2F	926909 > 926936	GTAAGGAAGAAAACTTATTCCACTATCG
WGAP22walk_3F	927530 > 927549	TCAACTGCCTCTGAATGTGC
WGAP22walk_4F	928127 > 928148	GGGTATGTAGGAAACTTGGTGG
WGAP22walk_4R	928760 > 928740	TGCTTTGTGTTATCGCCTACG
WGAP22walk_3R	929486 > 929462	TGGATGTATAATACTTAGGTGCTGC
WGAP22walk_1R	930111 > 930085	GATGTATTCACATGTTATCATGAGAGG
WGAP22_R	930592 > 930567	TGAACCTGGGTATTATAAAGATGGAG
WGAP22_2F	930700 > 930681	GCATGGTACAGGTCATGGTG

Primer name	Position	Primer sequence (5' > 3')
WTHIN568_4R	931810 > 931830	AAGATCCAACTCCTGTTGAGC
WTHIN568_4F	933035 > 933008	AACTTAATATTCGTTCCATACATAGCTG
WGAP21_F	936478 > 936498	TAGTTGCCCCAACTATAGCAC
WGAP21_R	937548 > 937522	AAGGTTTATGGACATTTAGCTTATAGG
WTHIN_940940_F	940928 > 940953	CTACCGATAACTGAAAACAATAACCC
WTHIN_942230_R	942218 > 942197	TCCGTCTTTTCCTGTATATCCC
WGAP20_F	943418 > 943441	TGTAACATTATCTCGCTGCATAGG
WGAP20_R	944607 > 944582	ATTATGTAACTGTTGCGTATTGTCAG
699_RCF	954033 > 954060	CATCAACCTATGTCTAGTGAGTATATCC
WTHIN568_3R	954533 > 954556	ACACCAAGATCTCATTCGATAACG
WTHIN10217_1R	955428 > 955454	TCAATTTTCAAATACCCAATAACTCCC
WTHIN568_3F	955888 > 955864	ACTCACGTATAACTGATCTATGCAG
WTHIN10217_1F	958420 > 958386	TGATATTAAACTATCTATCTTTTGTAAGTGTACTG
WGAP76_R	960661 > 960687	CACAATCATATGATTTTGAAACTCTGC
WGAP76walk_1R	960942 > 960970	AGCTAACAGTTTACTGAATAACATATCAC
WGAP76walk_1F	962014 > 961989	AGAACTCTAAATGATAATGTCTCGGG
6654_R	963134 > 963109	TGATGAGTAGATTCACAATAGTTCGG
WTHIN11030_3R	966121 > 966145	AAGTATATCGCAACATAATGTGAGC
25104_RCR	966217 > 966238	TTTCATAGCATCCTACCTTCCC
WGAP76_F	966242 > 966224	TAGGGGGAAGGTAGGATGC
11030_RI	966520 > 966492	GGTTTTGAGAGTAGTAGATATTTGTTAGG
WTHIN11030_3F	966660 > 966640	TGACCGAAGTGAATACCAAGC
WTHIN11030_2R	966978 > 967001	AGTTCCCATTCATTTAGTTTTGCC
WTHIN11030_2F	968032 > 968008	GTGAACATTTACTAGTGGTACATGC
WTHIN14308_2R	968270 > 968299	TCTCTAGAAAAGAGAATTACTGTTAGATCC
WTHIN14308_2F	968620 > 968597	ACAGTAAAGTTTGTAGAGGATGGG
WTHIN11030_1R	972270 > 972293	TTCACATTGAGTTACAATTCCAGC
WTHIN11030_1F	973098 > 973072	TCAGTTATAGAGTAAAGAGTTAGGACC
WTHIN_978928_F	978916 > 978943	CAATCTTAAGGTGATTATCTGTATAGCC
WTHIN_980451_R	980439 > 980414	AACAGTTATACTTCAAGATGACATGC
WTHIN14308_1R	980658 > 980686	GATAGATTATTCTTCGAAGATCCATTCAG
WTHIN14308_1F	981332 > 981307	GGGTGAACAAGAATTGAAATTATTGG
WGAP75_R	981789 > 981815	AGTCAATCATTATCAACCAACAAATCC
WTHIN12427_3R	982419 > 982441	CTCAATCATGTATCCGTCAACAC
WGAP75_F	982426 > 982398	TGATTGAGATATAGATGCTATAGTTGTCC
WTHIN12427_1R	982729 > 982757	CATGCTACTTTTACCACTATATACAAGAC
WTHIN12427_2R	983518 > 983543	CCAAAATACAATCTACGTATTCTCGC
WTHIN12427_2F	983744 > 983719	GTTGGTGCAGTAATAGTGTATAATGG
WTHIN12427_1F	984446 > 984416	TTATGACTTATGAGTCATATATTAAAGCCAG
WGAP74_R	1000142 > 1000163	TCTCCATAGCAGCATTATGCAC
WGAP74walk_1R	1000674 > 1000698	CTCCAGCTTCCTTAATTTTGATAGC
WGAP74walk_2R	1001178 > 1001204	GGAAGTACAGTTTTTATGGAATGTAGC
WTHIN_1001515_F	1001545 > 1001571	CAGAATTACAACGCTATATAGTAGCAG
WGAP74walk_3R	1001916 > 1001936	TCATTCCTGTGTTTCACATGC
WGAP74walk_4R	1002575 > 1002601	CATTATGTCCTTAGTACTCATAAGCAC
WTHIN_1002866_F	1002852 > 1002879	CTCACTTCAATCATCTAACAATATAGCC
WGAP74_F	1004365 > 1004344	GGGAAGTTTGTTTATCCACAGG
WGAP73_R	1005208 > 1005234	CTATAGCACGAAAATTATTATGACGCC
WTHIN_1005552_F	1005538 > 1005561	CCAATGCACTAACTACATTAGCAG
19566_F	1006580 > 1006555	TTGTAATAGAAAAAGTGGCTTTGGTG
WTHIN2177_1R	1021310 > 1021331	AAGCATGAGTCCTAACTACTGG
WTHIN2177_1F	1022226 > 1022204	AGTTACGTTGTTGCATTTAAGGG
2177_RCR	1022347 > 1022370	AGGTAACTGTAAGCTCATTAGTGC
2177_RCF	1023930 > 1023907	GGGTATAACTCAATGGTAGAGTGC
19566_R	1028526 > 1028549	CAGAAATTCAGCAGAGTTTAGACC
WGAP72_F	1031212 > 1031189	GGAGTGTAAGTATGGGTATAGCAG
WGAP71_R	1034459 > 1034484	CCATATTACTTCATCAAACACCTTGC
WGAP71walk_2F	1035673 > 1035648	ACAGTTCAAATGATTACCTTATGACG
WGAP71walk_1F	1036318 > 1036293	GAATTATGGTTTGTGTTATTGCATGC
WTHIN1575_1R	1036441 > 1036466	ACACTTAATAAAATCAACATCTGCCC

Primer name	Position	Primer sequence (5' > 3')
WGAP71_F	1037712 > 1037686	ATTGATGTGGAATATTGAGTACATAGC
WGAP70_R	1041036 > 1041062	TTCTCCTGTAATATCATCACTTAGCTC
WGAP70_F	1042166 > 1042139	GAGACAAAGTATATTTTGATGACCTGTG
WTHIN1403_1R	1043355 > 1043379	AAACAATCATTCAGATCTACAGCAC
WTHIN1403_1F	1044372 > 1044350	AGAAGTCATGAGTATCGTTACGG
WGAP69_3R	1044428 > 1044453	ACTATATGTAGCAATGACTAGAGCTG
WGAP69_2R	1044786 > 1044811	TTTATACAGTACAATTAATGCCGCTC
WGAP69_R	1044888 > 1044916	CTGTACCATAACTAGCTACATATAAGACC
WGAP69_F	1045897 > 1045872	TTTTGATTTGTCTATGACAGCTATGG
WTHIN1486_8R	1047049 > 1047076	ACTTAAACTCTGTTATCAGATTACTTGC
WTHIN1486_8F	1047916 > 1047892	AGTATATGATAGGTCAACCTTTGGG
WGAP68_R	1054385 > 1054406	ATAGTCATAGGACAAGGATGCG
WGAP68_F	1054602 > 1054580	ACTGTTGGAAGTATAGCACTTGG
WTHIN1486_7R	1057517 > 1057541	CAAGGAGTTCAAGTAATACCTATGC
WTHIN1486_7F	1059139 > 1059110	TTTTACTTGTTTCTAGTGTATAATATCCCG
WTHIN1486_6R	1062578 > 1062603	AACATCAATAATCAAGAAGTCATGGG
WTHIN1486_6F	1063428 > 1063403	TGTACATGTGATATATTGAGGGATGG
WGAP67_R	1064530 > 1064552	GCATGCCTGATATTTCAGAATCC
WGAP67_F	1064981 > 1064954	CAATGTTGCAAGTAAATCTATTAATGGC
WTHIN1486_5R	1065182 > 1065206	CTATAGGCTTGAAAATCACAAGACC
WTHIN1486_5F	1066325 > 1066306	TGTGACAAGACTGGTACAGG
WTHIN1486_4R	1067000 > 1067027	AACTTCAAATAACATTCATGGTATAGCG
WTHIN1486_4F	1067825 > 1067803	GAGGTGGATTTAATCTGTCTTGC
WTHIN10971_1R	1071558 > 1071585	CACTCTTATAGGTATAATACAACGTTGC
WTHIN10971_1F	1072938 > 1072913	CTTACGTTAATAGAAAAGGTGGTGTG
WTHIN1486_3R	1073040 > 1073065	ACATATCAGCTTCTAGTAAAACCACC
WTHIN1486_3F	1074680 > 1074658	TGACAGGAGAATTTATTGGGTGG
WGAP66_R	1077357 > 1077375	TGCAACAGCTGAAGTTTGC
WGAP66walk_1R	1077985 > 1078005	ACGCAATATTAGCACCTGTCG
WGAP66walk_1F	1078306 > 1078286	GCTTTGCTTGATGCATATGGG
WGAP66_F	1079440 > 1079415	TGATATGTATGAATCGTTTTTAGCCG
WGAP65_R	1085361 > 1085389	GAACTACCATTGTTCCTTATATTATCACC
19572_R	1085553 > 1085527	CTAGTTGTTACAACAATTTTGATTGGG
WGAP65_F	1086154 > 1086133	TAGCAGAAAAATTCATGGTGGC
WTHIN_1086186_R	1086175 > 1086151	AAGATCCTATGTGAGAAACACTAGC
19572_F	1110043 > 1110067	CTTCACAAGAAAGGCATAATACCTC
WGAP64walk_1R	1110432 > 1110458	ACTAATATGCTAAAACACACTTATGGG
WGAP64walk_1F	1111606 > 1111584	GGTGCCATCTTTGTTAGTAATGG
WGAP64_F	1111768 > 1111741	CCATGTATATAGTGTCATCTTTGAATGC
WTHIN1486_2R	1114274 > 1114299	CTGTTGCTACTGTTATAACTAAAGGC
758_RC1_F	1114712 > 1114732	AAGCAGTCAAATACCCATCCC
758_RC1_R	1115452 > 1115429	TCATGATGTAGTGAGAAAGTGTGC
WTHIN1486_1R	1115557 > 1115582	TCCTGACATAATAGATTCTATCACCC
WTHIN1486_2F	1115583 > 1115558	AGGGTGATAGAATCTATTATGTCAGG
WTHIN1486_1F	1116378 > 1116352	GTGTACTCATGTATTAAGTGATGAAGC
WTHIN_1125301_F	1125168 > 1125191	TCTCCATCAGAAATAGATAACGAC
WTHIN24731_6F	1125168 > 1125192	TCTCCATCAGAAATAGATAACGACC
WTHIN_1126360_R	1126435 > 1126410	GTATTGCTACGGTAATTTATCTGGTG
WTHIN_1130794_F	1130869 > 1130888	TGTTAGGACATCGTGGATGC
WTHIN_1131423_R	1131498 > 1131474	CCTTATTGAATATTGTGCATTGTGC
WGAP63_R	1134132 > 1134162	AGTATAGATTCTACATTCTACATACTACACC
17139_R	1134211 > 1134239	AAAATAGTCGTAAACTAGTAGCTTATCAC
WGAP63_1462R	1135280 > 1135252	TTACGATTAATATAATGGTGATGTATGGG
WTHIN_1136833_F	1136908 > 1136929	TTACTGGTTCAGTATGAGGCTC
WGAP63_1462F	1137707 > 1137732	CCTTGTACATCTTTCATATAAGCACC
WTHIN_1137750_F	1137825 > 1137804	GCAGGTCACTTATACAGTCTGG
1547_LE	1139197 > 1139169	ACTTTATAGCTAGTAGAGTGTGATATTGC
1547_RI	1141061 > 1141087	ACCAACAAAGAAATAGTTAACACTACC
896_LE	1141633 > 1141605	TTTAAGGTACATCTATTGTTTAAGGAGTG
WGAP63_F	1142673 > 1142648	CATAGTATTAGCGGTAGTACGTACAG

Primer name	Position	Primer sequence (5' > 3')
WTHIN13621_1R	1142854 > 1142883	TCCTTTAGATATTCAGTTACAACATGTCAC
WTHIN13621_1F	1144089 > 1144064	CCATAGATTCTAGTATCAGGCAATGC
WGAP62_R	1147598 > 1147623	ACACATAATTACACATCATTTGAGCG
WGAP62walk_1R	1148141 > 1148168	CTTACACTCTATAGTATATCCCGTTAGC
WTHIN24165_2R	1148782 > 1148806	AACATATTGTATGACAAAGCTCTGG
WGAP62walk_2R	1149141 > 1149169	TGTAATATCATACATATACTAACAGCGGG
WTHIN_1149078_F	1149185 > 1149211	TTTCACAATACTCAGTCACTTAATACC
WTHIN_1150475_R	1150550 > 1150525	CATATACCTTGTTCTTAAGGTAACGG
14311_R	1150656 > 1150634	AAACGATTTTTGGTGCTATTGGC
WTHIN24165_1R	1163877 > 1163894	CATCATCAGCTGCTTGCC
WTHIN24165_1F	1164141 > 1164114	GTTGGTATGTTATCAGTTATTCTCAAGG
14311_F	1170927 > 1170956	TCAAGGTAAACCAATACTACATATTACCTC
WGAP61walk_2R	1171088 > 1171112	ACATATCCACACTTACTAAACCCTC
WGAP61walk_2F	1172236 > 1172212	GTTATTAAGAGCATAGGCAAGTAGC
WGAP61walk_1F	1172343 > 1172318	ACTGGTATAATGATTATGAATGGGGG
WGAP61walk_3F	1172389 > 1172365	TGGTAACTCTATTTTGAAAGATGCG
WGAP61_F	1172904 > 1172878	ATGTAGGTAAATAGCTAAGGTGTATGG
WTHIN7701_1R	1174702 > 1174723	GCTTACGAACTGTAGAATTGGC
WTHIN7701_2R	1175366 > 1175390	TGGAGAATTATCAAGTATGTCACGG
WTHIN7701_1F	1176337 > 1176315	TGATGCAGGTATCATATTGGTGG
7702_LE	1186937 > 1186960	TTCTACTAGAGTGAGGGTTTATGC
5054_RI	1187351 > 1187328	TGAAATTTCTAGGGTCAGGATGTG
WTHIN5054_1R	1193363 > 1193389	TGATTACTATACAAATAACGTGTGGTG
WTHIN5054_1F	1194065 > 1194042	TGCATATTTCTGCAATGTGTACAG
5054_LE	1195073 > 1195101	ACACATAGATAAAGTTTTTGACACTATGC
WGAP121walk_1F	1195326 > 1195351	CAGCTAAAATAATTTAGCTTTGCACC
WGAP121walk_1R	1196279 > 1196246	TCTTAGATAGTTAATCAGTCTTTACAAATAATCC
7239_RI	1196509 > 1196489	GTTGCTCCAAAAGAGTTCCAG
WTHIN1674_2R	1201095 > 1201121	CTTAACAGTAGTTTCACAACTAGAAGC
WTHIN1674_2F	1202431 > 1202409	CTGACGTTCAAACTATAGAACGC
7239_LE	1206918 > 1206940	GCATGATGCCTTCAACTAATTCC
3850_RI	1207731 > 1207703	GTTACACACCATTACATATCATATAGAGC
WTHIN896_1F	1210444 > 1210467	TCTCACTATTCAATGGTCTATGGC
WTHIN896_1R	1210802 > 1210777	AAGAGAATATGCATAGATTGAGGGTG
WTHIN896_2F	1212816 > 1212843	AACCTATATATTCCATTCAATGCTTAGC
WTHIN896_2R	1213483 > 1213460	TGGTGAAGAAGATATTCAGTTGGG
3850_LE	1214106 > 1214134	ACCTATATTTTGATGAATCAGAAACACAC
WGAP188walk_1R	1214639 > 1214671	TCCTTTAGAGATCTATATTTATAATTCAAGCAC
8192_LE	1215853 > 1215828	TGTGTCTTATGAATTACATGATGCTC
8192_RI	1226711 > 1226739	ACAATTTCACAATATCTTAATACCAGCAC
4307_RI	1227652 > 1227631	GGAATACAAGTGATGATGGTGC
WGAP107_R	1231739 > 1231766	CTTACCAATTTCAAGTCTTAGTATGTCG
22799_R	1232239 > 1232213	TGTGATTGTGTTTGAAATACTAATGGC
WTHIN15591_14R	1233931 > 1233954	AACTTAAACACAATTACTGTGCCG
WTHIN15591_14F	1235634 > 1235611	GCCATTCCTAACTATGTACAGTG
WTHIN15591_13R	1240369 > 1240396	CAGCAATCATCAGATATATACTTCACAC
WTHIN15591_13F	1241467 > 1241441	ACAGTACCAATATGCTATTAAAAGTGC
WPCRJL1_check_1F	1259546 > 1259572	AACACACTGCTCATTATATATACATGC
WTHIN15591_12R	1259630 > 1259656	TCACATTAACATCAAAGAATTAGGCAG
WPCRJL1_check_1R	1260938 > 1260912	AGCACTACTAACAATAAACAAATACCC
WTHIN15591_12F	1261306 > 1261278	AATAACACAACTATAAGGTTAAAGTCTCG
WL2AP1_check_1F	1281261 > 1281282	ACATTCCACTTAAACTGCATGC
WL2AP1_check_1R	1282759 > 1282730	GGTATGTTTATCAGTTGTTAAGTACAAAGG
WTHIN15591_11R	1290476 > 1290499	ACAGTAACAGCAAATATGTGTAGC
WTHIN15591_11F	1292019 > 1291994	TTTTCGATAACTTTAGAATTGGGAGG
WTHIN15591_10R	1293749 > 1293775	CACTCATATATATCATGACATCAACGG
WTHIN15591_10F	1294969 > 1294946	CTCAGTCTTGTAAAATTGTCAGCG
WTHIN15591_9R	1295571 > 1295598	AGCAATTCATATTTATGCTACAGTATCC
22799_F	1296151 > 1296177	TTCCAACTTTTCTATCTGATTTTGCAG
WGAP48walk_1R	1296952 > 1296979	CACAAAACTAGGACTGTTAAAGTTAGAC

Primer name	Position	Primer sequence (5' > 3')
WTHIN15591_9F	1297159 > 1297132	TCATTAAGTGAAAGCTATTGTAATGCTC
WTHIN15591_8R	1297532 > 1297562	ACTACTACTACTTCAATATCAGTTAATCCAC
WGAP48walk_1F	1298437 > 1298408	TCAGATTATTAGTCGTAATATTATTGCTGC
WTHIN15591_8F	1299127 > 1299107	AGGCATTATCAAGAGGTGCAG
WGAP48_F	1299349 > 1299374	AAGTAGACAATAAATACTCGCTTTGC
WTHIN15591_15F	1299493 > 1299514	TCGTAACTTCCACAATACTCCC
WTHIN15591_15R	1300267 > 1300245	AGTGAACAAGAAATCTTGGATGC
WTHIN15591_7R	1300492 > 1300514	ACAACCCTATTGTACGATTACGC
WTHIN15591_6R	1301366 > 1301389	AATAACGTGCTTTTGGTCTAATCC
WTHIN15591_6F	1302260 > 1302235	TCATAATTTGTGGTTGAAAACGAGAG
WTHIN15591_5R	1306121 > 1306147	GATAACCCACTTAACCTGTATAATCAC
WTHIN15591_5F	1307235 > 1307209	GTGACATTAATGACATCAACCATAACC
WGAP47_R	1307594 > 1307623	TGAATCTAGTAACATATGTGATTTGTACAG
WGAP47_F	1308520 > 1308498	TTTGGTATTGTTGTGTGAAGCAG
WTHIN15591_4F	1308727 > 1308707	GTCTTTGTCACGACAAATGCC
WTHIN15591_3R	1312053 > 1312078	TTGTACAGTAGATCTGTACAATACCG
WTHIN15591_3F	1312865 > 1312842	GATCTTCTTGTTGAGTTACTTGGG
8028_RCR	1313586 > 1313606	TTGTATTGCCAGTTGTTGCAC
8028_RCF	1316525 > 1316499	ATGTGTGGATCTAGTAATTCATTAGTG
WGAP46_R	1320481 > 1320504	AGTAGACAGTATACAAGCGTTTCC
WGAP46_F	1320949 > 1320922	GGTAAGAATTTTGAGTGTAATTGTTGTC
WTHIN15591_2R	1325023 > 1325042	TGTTCGACACACTGACATCC
WGAP45_R	1325281 > 1325309	CAGCAACTATCAAATATACCAAAATTGAC
WGAP45_F	1326066 > 1326046	GAGTCTGAAAACCCATTTGCC
WTHIN24586_1F	1330242 > 1330269	CTAAATACTGAGAATTAGGGAACAAACG
WTHIN24586_1R	1330865 > 1330839	TTGTGAATTCTAGTATCAATCTTGTGG
WTHIN15591_1R	1331315 > 1331335	AGTAAAAGCAGATGGACTCGC
WTHIN15591_1F	1332066 > 1332044	CTAGATTTTCACCTTGTGCTACG
WGAP44_R	1334531 > 1334557	TGTATCAAATACAATTAGTAGCACCAC
WGAP104_F	1335112 > 1335086	CATGCTAATCTATGTGACAGTAAAGTG
WTHIN15591_17R	1336756 > 1336777	TCTGACTTAGCAGCAGATAACC
WTHIN15591_17F	1337040 > 1337020	ACAGGTATTCAATGGTGGAGC
WGAP43_F	1343829 > 1343852	GCAGATGATTATGTAACAAAGCCC
WGAP43_R	1344151 > 1344131	AGCTTTTTGCGCAACTTACAC
WGAP44_F	1347063 > 1347087	CATGCATTACACAGATCTTCAACAG
WTHIN2266_1F	1348121 > 1348096	CCACACTTTCAGTTCAATCTTATAGC
WGAP104_R	1348480 > 1348451	GAAATAGGTGTTTATTTCTGTTAGAAATGC
WGAP105_F	1352191 > 1352214	ACACGTTACCTACTCGTAAAACAC
WGAP105_R	1352756 > 1352730	TGTAAGTACAGATTGTGTAATTCAACG
WGAP106_F	1353905 > 1353931	ACTTCATTAATCTCATCATTAGATGGC
WGAP106_R	1354116 > 1354094	GCAGCAGTTTTATCTGGTAATCG
WTHIN2864_3R	1354685 > 1354664	TGGAATACTTGGAGATGACAGG
WTHIN2864_1F	1361969 > 1361993	CATCAGCATTAAAGTAACCTTGTCC
WTHIN2864_1R	1362667 > 1362646	CACAGCCTCTTAGATGTGTACC
WGAP113_F	1363692 > 1363717	CAATCATCATAGACTCAACTTACCAC
WGAP113_2R	1364243 > 1364219	TGAAATGTATACTCTAGACTGGGAC
WGAP113_R	1364783 > 1364755	GCTATTAAGTGGTATATGTGTTTATCAGG
WGAP54_F	1369934 > 1369962	TGACAATATATCACCTGACTATTAACCTC
WGAP54_R	1370929 > 1370906	TAGAGTTTGTATTGGGAGATCGTG
WTHIN2864_5R	1372623 > 1372598	GAAGTCATTAATTAGTCAATGGAGGG
WTHIN2864_6F	1376372 > 1376392	ACCTACCACAAACGCTATACC
WTHIN2864_6R	1377020 > 1376993	TGTTACCTAATAGTTAAGAATGATGTGC
WTHIN2864_7F	1381544 > 1381572	AATGTAGCAATACAATACTAAGAAAGAGG
WTHIN2864_7R	1382346 > 1382321	GATGTAATGAAGGAGAATTCTAACGG
WTHIN_1383570_F	1383645 > 1383665	AGTACCATTTACTCCACCTGC
WGAP9_F	1384020 > 1384049	CCAAATATTGTATAAACTCTACACTTTCTC
WGAP9_R	1385340 > 1385314	TTTGATAGTGTTGCTCGATATTATACC
WTHIN2864_9F	1392168 > 1392190	AATACCCTTCTTTCACAACAAGC
WGAP8_F	1392370 > 1392396	TTAACACATACCCAATAATACTGAAGG
WGAP8_R	1392900 > 1392872	CAGTTGAAAATTACTCTGTTAATTACTGG

Primer name	Position	Primer sequence (5' > 3')
WGAP7_F	1395735 > 1395763	ATTACGAAGCTATTACATTAGATGAAGTC
WGAP7_R	1396093 > 1396064	ATAGGGTGTATATATCTGTATAGATGGTAC
758_RC2_F	1396151 > 1396180	GATTATGTCAAAAACTAGTCATCTTTTAGC
758_RC2_R	1397602 > 1397578	GGATATAGTAAGTCAGTGTACTGGC
WGAP6_F	1402487 > 1402509	CTACTGGTATCTCCATACCAACC
WGAP6_R	1402652 > 1402634	CAATTCCAGCACCATTGGG
WTHIN2864_10F	1403051 > 1403071	ACAACAAAGCTGTTTTCACCC
WTHIN2864_10R	1404039 > 140401	CTGAATCAGAAGATGAAGAATTGTGC
WTHIN2864_11F	1414133 > 1414157	AGTGATATAGCAATTGATGGAATGG
WGAP5_F	1414867 > 1414894	ACATGATTGAGAATACCTTACAAACAAC
WGAP5_R	1415671 > 1415649	ATTGTTTGTGATGGGAATATGGG
WTHIN_1427957_F	1428033 > 1428056	AGACACGAATCAGGTAAATATCCC
WGAP4_F	1428287 > 1428311	GATGAAACCACAGAACAAAAAGGTC
WGAP4_R	1428556 > 1428537	AGATGCTCTCTGCATGTTGG
WGAP3_F	1438066 > 1438092	GAATATGATGCTATAATAGCTGCTCAC
WTHIN2864_14F	1439495 > 1439515	AGACCCAAGATCTACAACAGC
WGAP3_R	1439898 > 1439867	CATGGTAATAATTAAGATAATTTCTGGTGTTG
WTHIN2864_14R	1440686 > 1440658	GATGTTAATGGGAAGATTAGAGATTATGC
WTHIN2864_15F	1447197 > 1447221	AGAGACAGAAGACTCAGTATTTACC
WTHIN2864_15R	1448011 > 1447989	TCGCTGTGATAAAAGCTCTAAGG
WL2TP1_check_2F	1449621 > 1449649	TCTCAACTATAAACAATAGAGAATTTGGC
WL2TP1_check_2R	1450313 > 1450291	GCTAGACAAGTTTTGATGTTGGC
WL2TP1_check_1F	1450778 > 1450805	TCACCCTAAATTCTATCATAATTACAGC
WTHIN2864_16F	1450873 > 1450898	AGAACACTTGTTTACATATTGCTTGC
WTHIN2864_16R	1452277 > 1452252	GCAGGCATTAGATAATCTACAAGAGG
WGAP2_F	1458669 > 1458689	ATTTCCTCCAGTCGTAGTCTC
WGAP2_R	1458869 > 1458850	GCTGCATTAACACCTAGAGC
WGAP1_F	1463837 > 1463860	CTTAAACACTTACACCAATGCCAC
WGAP1_R	1464291 > 1464271	GGCCAGATGAATCTTCGACTG
WTHIN2864_18F	1466403 > 1466423	ACAACAAGCAAATGTAGCACC
WTHIN2864_18R	1466907 > 1466882	AGTGTTGCATGAAATACTTTGTATGG
WTHIN2864_19F	1471341 > 1471365	CTGAACCAATAGTAAAACTTGCAGC
WTHIN2864_19R	1472505 > 1472477	GAAATAGCATCATAAGAAAGTACTAGGTG
WTHIN2864_21F	1487529 > 1487552	CCCATTAGTAACATCTGCAATAGC
WTHIN2864_21R	1488337 > 1488311	GTCATGAACATATTAATGTGTGTCTGC
WTHIN2864_22F	1489822 > 1489842	GGTGGTGAACTTGGAGTAAGG
WTHIN2864_22R	1490915 > 1490890	TCAGAATTGATTGAGTAACTTATGCC
WTHIN2864_23F	1491063 > 1491087	CTTTGGTATAGAAATTGGAGGAAGG
WTHIN2864_23R	1492041 > 1492016	GTGTTTCCTGTTTAAGGATAAGATGG
WTHIN2864_24F	1492872 > 1492895	AGCATAGTACGAATACTTAGTGGG
WTHIN2864_24R	1493373 > 1493349	CTGGTTTAACTAAGGGTGTTTATGG
WGAP55_F	1498494 > 1498522	ACTGATAGAACATATAAACAACATCACCG
WGAP55walk_1F	1499373 > 1499395	AGTACACTGCTTTCTCATACCAC
WGAP55walk_1R	1500174 > 1500152	GTGAAATGTATGCGTAGTAGTTC
WGAP55_R	1500710 > 1500686	TGGACCTATAGAATTGGTTTACTGC
WGAP56_F	1505871 > 1505894	CAGAAAATGCAGAAGAAATTGCAC
WGAP56_R	1506892 > 1506868	GTATTTCTATAATGTCGCATGACCC
WTHIN2864_26F	1507985 > 1508005	GTTTGGTAAAGTTCGCAGAGC
WTHIN2864_26R	1509300 > 1509276	CGATTTCTATTACAGGATTTGCAGG
WGAP57_F	1512265 > 1512287	GAAAATCTACCAAAACGAGAGGG
WGAP57_R	1513153 > 1513122	ATGCTTATTATAGTGATAATGTTGGATATCAG
WGAP58_F	1514697 > 1514716	ATAGATGCCCCCATCAAAGC
WGAP58_R	1515427 > 1515404	AACATGTGGTGCAATTTATACAGG

Duplicates	
WGAP48R (= 22799_R)	24993_557_F (= WTHIN440_5F)
WGAP35F (= 16700_F)	24993_557_R (= WTHIN440_5R)
17139_F (= WGAP34F)	WL2TP1_check_1R (= WTHIN2864_16R)
3687_RI (= WGAP49_2F)	WTHIN1674_16R (= WTHIN1174_1R)

C2: Primers designed to amplify the *tuf* and *rho* regions.

Primer name	Primer sequence (5' > 3')
tuf_1	GCAAACAGGTGGTGCTGG
tuf_2	CATTTTCTTGCGCATAGACTCC
tuf_3	CCAGGATCTTGACACTGACC
tuf_4	TCCATAACACCAATATCCTGC
rho_1	ACACCTGTTGCACGTCG
rho_2	ACAAAGCAAGCCATGAAGC
rho_3	GACAACCTGAACATGCTCC
rho_4	CCTATCCATTCTCCAATCTTTTGC

C3: Primers designed to amplify and clone ORFs into the pCMViUBs and TOPO® pET vectors. Restriction enzyme sites are underlined.

Primer name	Primer sequence (5' > 3')
3630F_BamHI	GGATCCTTTACATTACAAAAACAATTTAACAGTAC
3630R_SalI	GTCGACTTACACTGCATGCCCT
4470F_BamHI	GGATCCAATGATTTCTCATTGTCTGGT
4470R_SalI	GTCGACTTAAAACTTAAACTTTGTACCTATCAA
5270F_BamHI	GGATCCATGTTTACTTTGCCAGAACTG
5270R_SalI	GTCGACTTATTTAACATTATCAATACATTGAGAA
5400F_BamHI	GGATCCATGCAAAACGTAAAATATATATTGTTTTG
5400R_SalI	GTCGACTCAATAAGTATTTAATACTAATGTATTACC
5430F_BamHI	GGATCCATGTTTGAATCTTTAACTAGTAGTTTAAC
5430R_SalI	GTCGACTTATTCATTGTTTTTCAGTAAATTCAT
7300F_BamHI	GGATCCATGAATCAGCAAATGGTAGTG
7300R_SalI	GTCGACTCACTCATGATTAACACCAC
8050F_BamHI	GGATCCTCTGAAGATATTGAGCAATATGATC
8050R_SalI	GTCGACTTACTTCTTTAACTTAACAGGAATAAATATTG
pET3630F	CACCTTTTTTACATTACAAAAACAATTTAACAGTACAAC
pET3630R	CACTGCATGCCCTATGTAAC
pET4470F	CACCTAATGATTTCTCATTGTCTGGTAAT
pET4470R	AAACTTAAACTTTGTACCTATCAAA
pET5270F	CACCATGTTTACTTTGCCAGAACTG
pET5270R	TTTAACATTATCAATACATTGAGAAAATC
pET5400F	CACCATGCAAAACGTAAAATATATATTGTTTTGG
pET5400R	ATAAGTATTTAATACTAATGTATTACCACTC
pET5430F	CACCATGTTTGAATCTTTAACTAGTAGTTTAAC
pET5430R	TTCATTGTTTTTCAGTAAATTCATAAAAT
pET7300F	CACCATGCCTGAGCAAAATGTATC
pET7300R	TAACTCATGATTAACACCACGTGC

C4: Vector specific primers used in this study.

Primer name	Primer sequence (5' > 3')
pET TrxFus Forward	TTCCTCGACGCTAACCTG
pET T7 Reverse	TAGTTATTGCTCAGCGGTGG
pCMViUBs IECO	GGCTAGCCTCGAGAATTC
pCMViUBs CMV991	CAGGGATGCCACCCGGG
pGEM SP6	ATTTAGGTGACACTATAG
pGEM T7	TAATACGACTCACTATAGGG

Appendix D: Protein classification scheme

0.0.0 Unknown function, no known homologs
0.0.1 Conserved in Rickettsiales
0.0.2 Conserved in organism other than Rickettsiales

1.0.0 Cell processes
- 1.1.1 Chemotaxis and mobility
- 1.2.1 Chromosome replication
- 1.3.1 Chaperones

1.4.0 Protection responses
- 1.4.1 Cell killing
- 1.4.2 Detoxification
- 1.4.3 Drug/analog sensitivity
- 1.4.4 Radiation sensitivity

1.5.0 Transport/binding proteins
- 1.5.1 Amino acids and amines
- 1.5.2 Cations
- 1.5.3 Carbohydrates, organic acids, alcohols
- 1.5.4 Anions
- 1.5.5 Other
- 1.5.6 Type IV secretion
- 1.5.7 ABC transporters

1.6.0 Adaptation
- 1.6.1 Adaptations, atypical conditions
- 1.6.2 Osmotic adaptation
- 1.6.3 Fe storage

1.7.1 Cell division

2.0.0 Macromolecule metabolism

2.1.0 Macromolecule degradation
- 2.1.1 Degradation of DNA
- 2.1.2 Degradation of RNA
- 2.1.3 Degradation of polysaccharides
- 2.1.4 Degradation of proteins, peptides, glycoproteins

2.2.0 Macromolecule synthesis, modification
- 2.2.01 Amino acyl tRNA synthesis, tRNA modification
- 2.2.02 Basic proteins - synthesis, modification
- 2.2.03 DNA - replication, repair, modification
- 2.2.04 Glycoprotein
- 2.2.05 Lipopolysaccharide
- 2.2.06 Lipoprotein
- 2.2.07 Phospholipids
- 2.2.08 Polysaccharides - (cytoplasmic)
- 2.2.09 Protein modification
- 2.2.10 Proteins, translation and modification
- 2.2.11 RNA synthesis, modification, DNA transcription
- 2.2.12 tRNA

3.0.0 Metabolism of small molecules

3.1.0 Amino acid biosynthesis
- 3.1.01 Alanine
- 3.1.02 Arginine
- 3.1.03 Asparagine
- 3.1.04 Aspartate
- 3.1.05 Chorismate
- 3.1.06 Cysteine
- 3.1.07 Glutamate
- 3.1.08 Glutamine
- 3.1.09 Glycine
- 3.1.10 Histidine
- 3.1.11 Isoleucine
- 3.1.12 Leucine
- 3.1.13 Lysine
- 3.1.14 Methionine
- 3.1.15 Phenylalanine
- 3.1.16 Proline
- 3.1.17 Serine
- 3.1.18 Threonine
- 3.1.19 Tryptophan
- 3.1.20 Tyrosine
- 3.1.21 Valine

3.2.0 Biosynthesis of cofactors, carriers
- 3.2.01 Acyl carrier protein (ACP)
- 3.2.02 Biotin
- 3.2.03 Cobalamin
- 3.2.04 Enterochelin
- 3.2.05 Folic acid
- 3.2.06 Heme, porphyrin
- 3.2.07 Lipoate
- 3.2.08 Menaquinone, ubiquinone
- 3.2.09 Molybdopterin
- 3.2.10 Pantothenate
- 3.2.11 Pyridine nucleotide
- 3.2.12 Pyridoxine
- 3.2.13 Riboflavin
- 3.2.14 Thiamin
- 3.2.15 Thioredoxin, glutaredoxin, glutathione
- 3.2.16 Biotin carboxyl carrier protein (BCCP)

3.3.0 Central intermediary metabolism
- 3.3.01 2'-Deoxyribonucleotide metabolism
- 3.3.02 Amino sugars
- 3.3.03 Entner-Douderoff
- 3.3.04 Gluconeogenesis
- 3.3.05 Glyoxylate bypass
- 3.3.06 Incorporation metal ions
- 3.3.07 Miscellaneous glucose metabolism
- 3.3.08 Miscellaneous glycerol metabolism
- 3.3.09 Non-oxidative branch, pentose pathway
- 3.3.10 Nucleotide hydrolysis
- 3.3.11 Nucleotide interconversions
- 3.3.12 Oligosaccharides
- 3.3.13 Phosphorus compounds
- 3.3.14 Polyamine biosynthesis
- 3.3.15 Pool, multipurpose conversions of intermediary metabolism
- 3.3.16 S-adenosyl methionine
- 3.3.17 Salvage of nucleosides and nucleotides
- 3.3.18 Sugar-nucleotide biosynthesis, conversions
- 3.3.19 Sulfur metabolism
- 3.3.20 amino acids
- 3.3.00 other

3.4.0 Degradation of small molecules
- 3.4.1 Amines
- 3.4.2 Amino acids
- 3.4.3 Carbon compounds
- 3.4.4 Fatty acids
- 3.4.5 Other

3.5.0 Energy metabolism, carbon
- 3.5.1 Aerobic respiration
- 3.5.2 Anaerobic respiration
- 3.5.3 Electron transport
- 3.5.4 Fermentation
- 3.5.5 Glycolysis
- 3.5.6 Oxidative branch, pentose pathway
- 3.5.7 Pyruvate dehydrogenase
- 3.5.8 TCA cycle

3.6.0 Fatty acid biosynthesis
3.6.1 Fatty acid and phosphatidic acid biosynthesis

3.7.0 Nucleotide biosynthesis
- 3.7.1 Purine ribonucleotide biosynthesis
- 3.7.2 Pyrimidine ribonucleotide biosynthesis

4.0.0 Cell envelope
- 4.1.0 Periplasmic/exported/lipoproteins
- 4.1.1 Inner membrane
- 4.1.2 Murein sacculus, peptidoglycan
- 4.1.3 Outer membrane constituents
- 4.1.4 Surface polysaccharides & antigens
- 4.1.5 Surface structures

4.2.0 Ribosome constituents
- 4.2.1 Ribosomal and stable RNAs
- 4.2.2 Ribosomal proteins - synthesis, modification
- 4.2.3 Ribosomes - maturation and modification

5.1.0 Laterally acquirred elements
- 5.1.1 Colicin-related functions
- 5.1.2 Phage-related functions and prophages
- 5.1.3 Plasmid-related functions
- 5.1.4 Transposon/insertion element-related functions

6.0.0 Regulation
- 6.1.1 Global regulatory functions

7.0.0 Not classified (included putative assignments)
7.1.1 DNA sites, no gene product
7.2.1 Cryptic genes

Appendix E: *E. ruminantium* gene list

The first column indicates the systematic identification number of each predicted ORF, followed by the gene name, protein product and length in amino acids. Columns 5 to 7 show the transmembrane helices and signal sequences predited by TMHMM2.0, SignalP3.0 and Phobius (th = transmembrane helix). Columns 8 and 9 represent the subcellular localisation predictions by CELLO and pSORTb2.0: C = cytoplasmic, P = periplasmic, IM = inner membrane, OM = outer membrane, E = extra cellular, U = unknown. In column 10 helix-turn-helix motifs are represented by plus signs. The size and frequency of tandem repeats, the EC number and functional class are given in the last three columns.

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
0010	*gapB*	NAD(P)-dependent glyceraldehyde 3-phosphate dehydrogenase†	335				C	C			1.2.1.59	3.3.15
0020	*elbB*	enhancing lycopene biosynthesis protein 2*	220				C	C				7.0.0
0030	*proC*	pyrroline-5-carboxylate reductase	271				OM	U			1.5.1.2	3.1.16
0040	*dnaZ*	DNA polymerase III, gamma subunit*	487				C	U			2.7.7.7	2.2.03
0050		unknown	101				C	U				7.0.0
0060	*asd*	aspartate-semialdehyde dehydrogenase	337				C	U			1.2.1.11	3.1.0
0070	*metK*	S-adenosylmethionine synthetase	401				C	C			2.5.1.6	3.3.16
0080	*ubiF*	2-octaprenyl-3-methyl-6-methoxy-1,4-benzoquinol hydroxylase*	390	1			C	U		158 bp x 2.8 (C-terminus)	1.14.13.-	3.2.08
0090		membrane protein*	193	1	signal	2 th	P	U				4.1.1
0110	*glyQ*	glycyl-tRNA synthetase alpha chain	280				C	C			6.1.1.14	2.2.01
0120	*glyS*	glycyl-tRNA synthetase beta chain	702				C	U			6.1.1.14	2.2.01
0130	*dnaJ*	chaperone protein DnaJ	382				OM	C				1.3.1
0140	*nadC*	nicotinate-nucleotide pyrophosphorylase [carboxylating]	277				C	C			2.4.2.19	3.2.11
0150		integral membrane protein*	195	5		5 th	IM	IM				4.1.1
0160	*ruvC*	crossover junction endodeoxyribonuclease RuvC	160				C	U			3.1.22.4	2.2.03
0170	*coxC*	cytochrome c oxidase subunit III	274	7		7 th	IM	IM			1.9.3.1	3.5.3
0180	*hemE*	uroporphyrinogen decarboxylase	335				C	U			4.1.1.37	3.2.06
0190	*corC*	magnesium and cobalt efflux protein†	288			1 th	C	C				1.5.2
0200		protease†	178				C	U				5.1.2
0210		genetic exchange protein†	394				E	U				5.1.2
0220	*bioC*	biotin synthesis protein BioC†	249				OM	IM				3.2.02
0230	*nadA*	quinolinate synthetase A	314				C	C			1.4.3.-	3.2.11
0240	*fdxA*	ferredoxin	125				C	C				3.5.3
0250		unknown	457				C	C		297 bp x 2.8		0.0.0

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
0260	*virD4*	type IV secretion system protein VirD4	801	3		2 th	C	IM		6 bp x 5.0		1.5.6
0270	*virB11*	type IV secretion system protein VirB11	332				C	U				1.5.6
0280	*virB10*	type IV secretion system protein VirB10	448	1		1 th	P	U		6 bp x 9.0		1.5.6
0290	*virB9*	type IV secretion system protein VirB9	267		signal	signal	E	U				1.5.6
0300	*virB8*	type IV secretion system protein VirB8	232	1		1 th	OM	U				1.5.6
0310		riboflavin biosynthesis protein*	371				C	C			3.5.4.25	3.2.13
0320		unknown	354				C	U				7.0.0
0330		integral membrane protein*	159	2‡		2 th	C	C				4.1.1
0340	*dapF*	diaminopimelate epimerase	265				C	C			5.1.1.7	3.1.13
0350		Unknown	143				C	U				0.0.0
0360	*pgk*	phosphoglycerate kinase	395				C	C			2.7.2.3	3.3.15
0370	*xseA*	exodeoxyribonuclease VII large subunit	388				C	U		203 bp x 3.0 (N-terminus)	3.1.11.6	2.1.1
0380		membrane protein*	222	1		1 th	C	U		203 bp x 3.0 (C-terminus)		4.1.1
0390	*dapD*	2,3,4,5-tetrahydropyridine-2,6-dicarboxylate N-succinyltransferase	284				C	IM			2.3.1.117	3.1.13
0400	*trmE*	tRNA modification GTPase*	439				C	C				2.2.12
0410	*dfp*	DNA/pantothenate metabolism flavoprotein*	181			signal	C	U			4.1.1.36, 6.3.2.5	2.2.03
0420	*recG*	ATP-dependent DNA helicase RecG	677			2 th	C	IM			3.6.1.-	2.2.03
0430		NADH-ubiquinone oxidoreductase subunit†	320	2		1 th	C	IM		283 bp x 3.2 (C-terminus)	1.6.99.3	3.5.3
0440	*dksA*	DnaK suppressor protein*	151				C	C				1.6.1
0450	*ccmB*	heme exporter protein B†	220	5		6 th	IM	IM				1.5.7
0460		cation efflux system protein*	306	6		6 th	IM	IM				1.5.2
0470		exported protein*	208		signal	signal	C	C				4.1.0
0480	*rpsT*	30S ribosomal protein S20	95			signal	C	U				4.2.2
0490	*polA*	DNA polymerase I	865				OM	C	+	198 bp x 3.0 (C-terminus)	2.7.7.7	2.2.03
0500		unknown	102				C	U				0.0.0
0510	*argF*	ornithine carbamoyltransferase	305				C	C			2.1.3.3	3.1.02
0520	*recF*	DNA replication and repair protein RecF*	372				OM	C				2.2.03
0530		uracil DNA glycosylase†	263			2 th	C	U			3.2.2.-	2.2.03
0540	*def1*	peptide deformylase 1*	181				C	C			3.5.1.88	2.2.09
0550	*plsC*	1-acyl-sn-glycerol-3-phosphate acyltransferase*	241	3		2 th	IM	IM			2.3.1.51	3.6.1
0560	*rpe*	ribulose-phosphate 3-epimerase	215				C	U			5.1.3.1	3.3.09
0570		integral membrane protein*	265	6		7 th	IM	IM				4.1.1
0580		ABC transporter, ATP binding protein*	239				C	U				1.5.7
0590		integral membrane protein*	613	3		3 th	OM	U				4.1.1
0600	*ispB*	octaprenyl-diphosphate synthase	325				C	C			2.5.1.-	3.2.08

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
0610	*glnA*	glutamine synthetase	470				OM	C			6.3.1.2	3.1.08
0620	*tyrS*	tyrosyl-tRNA synthetase	418				C	C			6.1.1.1	2.2.01
0630	*hemA*	5-aminolevulinic acid synthase	398				C	C			2.3.1.37	3.2.06
0631		conserved hypothetical protein	122			2 th	C	IM				0.0.2
0640	*secF*	protein-export membrane protein SecF	289	6		6 th	IM	IM				1.5.5
0650	*fbaB*	fructose-bisphosphate aldolase class I*	300				C	C			4.1.2.13	3.3.04
0660		unknown	3715				OM	U		300 bp x 2.6, 171 bp x 2.4, 471 bp x 2.7, 171 bp x 2.4		0.0.0
0670	*pdhC*	dihydrolipoamide acetyltransferase, E2 component of pyruvate dehydrogenase complex	406				C	C			2.3.1.12	3.5.7
0680		unknown	540				OM	U				0.0.0
0690		unknown	470			1 th	OM	U				0.0.0
0700		integral membrane protein*	547	2		2 th, signal	OM	U				4.1.1
0710		unknown	123			signal	C	U				0.0.0
0720		unknown	931			1 th, signal	OM	U				0.0.0
0730		unknown	93				C	U				0.0.0
0740	*guaA*	GMP synthase [glutamine-hydrolyzing]	528				C	U		170 bp x 4.0 (C-terminus)	6.3.5.2	3.7.1
0750	*gltA*	citrate synthase	415				C	C			2.3.3.1	3.5.8
0770	*gshA*	gamma-glutamylcysteine synthetase†	399				OM	U				3.2.18
0780	*valS*	valyl-tRNA synthetase	810				C	C		336 bp x 2.9 (C-terminus)	6.1.1.9	2.2.01
0790	*smpB*	SsrA-binding protein	148				C	U				2.2.10
0800	*ribB*	3,4-dihydroxy-2-butanone 4-phosphate synthase	211				C	C				3.2.13
0810	*greA*	transcription elongation factor GreA	162				C	U				2.2.11
0820	*atpA*	ATP synthase alpha chain	507				OM	C			3.6.3.14	3.5.9
0830	*atpH*	ATP synthase delta chain*	189				OM	U			3.6.3.14	3.5.9
0831		integral membrane protein*	84	2		2 th	IM	U				4.1.1
0840		integral membrane protein*	413	2		1 th, signal	C	C				4.1.1
0850		membrane protein*	258	1		1 th	C	C	+			4.1.1
0860	*lolE*	lipoprotein releasing system transmembrane protein LolE*	411	4		4 th	IM	IM	+			1.5.7
0870		conserved hypothetical protein	339				C	U				0.0.1
0880	*ccmF*	cytochrome c-type biogenesis protein CcmF	638	14		15 th	IM	IM				2.2.13
0890		aminomethyl transferase*	280				OM	U			2.1.2.10	3.3.00
0900	*purF*	glutamine phosphoribosylpyrophosphate amidotransferase	466				C	U			2.4.2.14	3.7.1
0910	*pth*	peptidyl-tRNA hydrolase	193				C	U			3.1.1.29	2.2.01
0920	*rplY*	50S ribosomal protein L25*	208				C	U				4.2.2
0930	*comF*	competence protein F†	230				C	U				7.0.0
0940	*dapE*	succinyl-diaminopimelate desuccinylase*	383				E	C			3.5.1.18	3.1.13

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
0950		glutathione-regulated potassium-efflux system protein*	569	10		13 th	IM	IM				1.5.2
0960		conserved hypothetical protein	193				P	U				0.0.2
0970		integral membrane protein*	155	4		4 th	IM	IM				4.1.1
0980	*pdhB*	pyruvate dehydrogenase E1 component, beta subunit*	332				C	C			1.2.4.1	3.5.7
0990		integral membrane protein*	607	2		2 th	OM	OM				4.1.1
1000	*tldD*	TldD protein	475				OM	C				6.0.0
1010		conserved hypothetical GTP-binding protein	363				C	C				0.0.2
1020	*ispF*	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase	173				C	U		154 bp x 5.1 (C-terminus)	4.6.1.12	3.2.08
1030	*ispD*	2-C-methyl-D-erythritol 4-phosphate cytidylyltransferase*	242			signal	C	U			2.7.7.60	3.2.08
1040		integral membrane protein*	1165	2		2 th	E	C		294 bp x 2.8		4.1.1
1050		integral membrane protein*	454	2		2 th	OM	U				4.1.1
1060	*purE*	phosphoribosylaminoimidazole carboxylase catalytic subunit	170				E	U			4.1.1.21	3.7.1
1070		exported protein*	180		signal	1 th	C	U				4.1.0
1080	*ihfA*	integration host factor alpha-subunit*	99				C	C				2.2.03
1090		conserved hypothetical protein	122				C	U	+			0.0.2
1100		unknown	161				E	U				0.0.0
1110		unknown	661				E	C		2 bp x 2.5, 27 bp x 56.0		0.0.0
1120	*trpS*	tryptophanyl-tRNA synthetase	332				OM	C			6.1.1.2	2.2.01
1130	*grpE*	GrpE protein	199				C	C				1.3.1
1140	*ribD*	riboflavin biosynthesis protein RibD	365				C	C			3.5.4.26, 1.1.1.193	3.2.13
1150		unknown	179				CP	C				0.0.0
1160	*pyrG*	CTP synthase	540			signal	C	U			6.3.4.2	3.7.2
1170	*secG*	protein-export membrane protein SecG*	110	2		1 th, signal	P	U				1.5.0
1180		integrase/recombinase XerD or XerC*	312				C	C				2.2.03
1190	*lolD*	lipoprotein releasing system ATP-binding protein LolD	228			signal	C	IM				1.5.7
1200	*maeB*	NADP-dependent malic enzyme	755				IM	IM			1.1.1.40	3.3.15
1210		exported protein*	877		signal	signal	OM	OM				4.1.0
1220	*lnt*	apolipoprotein N-acyltransferase*	506	7‡		7 th	IM	IM			2.3.1.-	2.2.06
1230		unknown	186				C	IM		237 bp x 2.4		0.0.0
1240		NADH-quinone oxidoreductase subunit*	492	13		14 th	IM	IM			1.6.99.5	3.5.3
1250		membrane protein*	99	1		1 th	P	U				4.1.1
1260		membrane protein*	149	1‡		1 th	C	U				4.1.1
1270		unknown	93				P	U				7.0.0
1280		conserved hypothetical protein	153				C	U				0.0.2
1290		unknown	564				E	U				0.0.0
1300		unknown	1334				E	OM				7.0.0
1310	*fbpA*	iron-binding periplasmic protein*	348		signal	signal	C	P				1.5.2

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
1320	*rpsP*	30S ribosomal protein S16	87				C	U				4.2.2
1330	*proP*	proline/betaine transporter	422	12		12 th	IM	IM				1.5.1
1340		conserved hypothetical protein	55				C	U				0.0.2
1350		short chain dehydrogenase*	231				C	U			1.1.1.-	7.0.0
1360	*pheS*	phenylalanyl-tRNA synthetase alpha chain	344				C	C			6.1.1.20	2.2.01
1370	*rplT*	50S ribosomal subunit protein L20	123				C	U				4.2.2
1380	*rpmI*	50S ribosomal protein L35	66				C	U				4.2.2
1390		conserved hypothetical protein	221				C	U				0.0.1
1400	*rho1*	transcription termination factor 1	478				C	C				2.2.11
1410		unknown	80				C	U				7.0.0
1420		dihydrolipoamide dehydrogenase, E3 component of pyruvate or 2-oxoglutarate dehydrogenase complex*	469				C	C			1.8.1.4	3.5.0
1430		unknown	951			1 th	OM	U		198 bp x 2.3		0.0.0
1440		membrane protein*	482	1		1 th	C	IM				4.1.1
1450		membrane protein*	208	1		1 th	OM	U				4.1.1
1460		exported protein*	180		signal	signal	C	C				4.1.0
1470		unknown	262				C	U				7.0.0
1480		truncated glutamine synthetase†	268				C	C			6.3.1.2	3.1.08
1490		ABC transporter, membrane-spanning protein†	374	8		9 th	IM	IM				1.5.7
1500	*alaS*	alanyl-tRNA synthetase	887				C	C			6.1.1.7	2.2.01
1510	*sucD*	succinyl-CoA synthetase, alpha subunit	295				C	U			6.2.1.5	3.5.8
1520	*sucC*	succinyl-CoA synthetase, beta subunit	386				C	U			6.2.1.5	3.5.8
1530	*rpsU*	30S ribosomal protein S21†	112	1		1 th	C	U				4.2.2
1540		exported protein*	342		signal	signal	C	U				4.1.0
1550	*map2*	major antigenic protein 2	209	1‡		1 th	P	U				4.1.3
1560		2-nitropropane dioxygenase*	345				C	U				3.3.0
1570		cytochrome b561*	173	5‡		5 th	IM	IM				3.5.3
1580		ABC transporter, membrane-spanning protein*	536	12		13 th	IM	IM				1.5.7
1590		secretion protein*	514	1		2 th	OM	IM				1.5.5
1600		unknown	204				C	U				0.0.0
1610		conserved hypothetical protein	542				C	U				0.0.2
1620		integral membrane protein*	197	4		4 th	IM	IM				4.1.1
1630	*rpsL*	30S ribosomal protein S12	123				P	U				4.2.2
1640	*rpsG*	30S ribosomal protein S7	160				C	U				4.2.2
1650	*fusA*	elongation factor G	689				C	C				2.2.10
1660	*tufA*	elongation factor Tu-A	395				C	C				2.2.10
1670	*nusG*	transcription antitermination protein NusG	179				C	C				2.2.11
1680	*rplK*	50S ribosomal protein L11	147				P	U				4.2.2
1690	*rplA*	50S ribosomal protein L1	220				C	U				4.2.2

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
1700	*rplJ*	50S ribosomal protein L10	160				C	IM				4.2.2
1710	*rplL*	50S ribosomal protein L7/L12	131				C	C				4.2.2
1720	*rpoB*	DNA-directed RNA polymerase beta chain	1380				C	C			2.7.7.6	2.2.11
1730	*rpoC*	DNA-directed RNA polymerase beta' chain	1411				C	C			2.7.7.6	2.2.11
1740	*bioF*	8-amino-7-oxononanoate synthase*	367				C	IM			2.3.1.47	3.2.02
1750		integral membrane protein*	142	4		4 th	IM	IM				4.1.1
1760	*rnhB*	ribonuclease HII	212				C	U		208 bp x 3.6 (C-terminus)	3.1.26.4	2.1.2
1770		unknown	1529				OM	OM				0.0.0
1780		Na+/H+ antiporter subunit†	172	2		2 th	IM	IM				1.5.2
1790		membrane protein*	205	1		1 th	C	U				4.1.1
1800		unknown	257				IM	U				7.0.0
1810	*pyrD*	dihydroorotate dehydrogenase	346				OM	U			1.3.3.1	3.7.2
1820	*def2*	peptide deformylase 2*	194				C	C			3.5.1.88	2.2.09
1830	*argH*	argininosuccinate lyase	462				C	U		202 bp x 2.0 (N-terminus)	4.3.2.1	3.1.02
1840		unknown	267				C	U				7.0.0
1850	*pdxH*	pyridoxamine 5'-phosphatate oxidase	194				C	U			1.4.3.5	3.2.12
1851		unknown	92				C	C				0.0.0
1860		membrane protein*	270	1‡		signal	C	U				4.1.1
1870	*dnaE*	DNA polymerase III, alpha subunit	1119				C	C			2.7.7.7	2.2.03
1880	*aroE*	3-phosphoshikimate 1-carboxyvinyltransferase	427			1 th	IM	IM			2.5.1.19	3.1.05
1890	*sdhC*	succinate dehydrogenase cytochrome b-556 subunit*	132	3		3 th	IM	IM			1.3.5.1	3.5.3
1891	*sdhD*	succinate dehydrogenase cytochrome b small subunit*	116	3		3 th	IM	IM			1.3.5.1	3.5.3
1900		unknown	417				OM	U				0.0.0
1910	*thiD*	phosphomethylpyrimidine kinase*	266				C	U			2.7.4.7	3.2.14
1920		conserved hypothetical protein	230				C	U				0.0.2
1930		integral membrane protein*	373	8		7 th	IM	IM				4.1.1
1940	*rpsD*	30S ribosomal protein S4	202				P	U				4.2.2
1950		conserved hypothetical protein	69				C	C				0.0.2
1960		exported protein*	383		signal	signal	OM	U				4.1.0
1970		acetyltransferase†	262				C	U			2.3.1.-	7.0.0
1980	*pgpA*	phosphatidylglycerophosphatase A*	168	4		4 th	IM	IM			3.1.3.27	2.2.07
1990	*tig*	trigger factor	446				C	U				1.3.1
2000	*clpP*	ATP-dependent Clp protease proteolytic subunit	198				C	U			3.4.21.92	2.1.4
2010	*clpX*	ATP-dependent Clp protease ATP-binding subunit ClpX	406				C	C				2.1.4
2020	*lon*	ATP-dependent protease La	801			signal	C	C			3.4.21.53	2.1.4
2030	*fmt*	methionyl-tRNA formyltransferase	303				C	U			2.1.2.9	2.2.01
2040		conserved hypothetical protein	272				C	U				0.0.2

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
2050		conserved hypothetical protein	285				C	C				0.0.2
2060	*thiE*	thiamine-phosphate pyrophosphorylase*	350				C	U			2.5.1.3	3.2.14
2070		integral membrane protein*	431	4‡		2 th, signal	IM	IM				4.1.1
2080		integral membrane protein*	94	2		3 th	IM	U				4.1.1
2090	*ftsK*	cell division protein FtsK*	855	5‡		4 th	IM	IM		45 bp x 4.1		1.7.1
2100		integral membrane protein*	307	6		6 th	IM	IM				4.1.1
2110	*argD*	acetylornithine/succinyldiaminopimelate aminotransferase	391				C	U			2.6.1.11, 2.6.1.17	3.1.0
2120		histidine kinase sensor component of a two-component regulatory system†	477	2‡		1 th, signal	OM	IM				6.1.2
2130	*mutL*	DNA mismatch repair protein MutL	689				C	U				2.2.03
2140	*smf*	DNA processing protein chain A*	375				C	U				2.2.03
2150	*fabF*	3-oxoacyl-[acyl-carrier-protein] synthase II	423				C	IM			2.3.1.41	3.6.1
2160	*acpP*	acyl carrier protein	92				C	C				3.2.01
2170		unknown	1073				E	OM		252 bp x 2.7		0.0.0
2180		integral membrane protein*	876	2		2 th	OM	C				4.1.1
2190	*rpmG*	50S ribosomal protein L33	56			1 th	C	U				4.2.2
2200		integral membrane protein*	235	6		6 th	IM	IM				4.1.1
2210	*dsbB*	disulfide bond formation protein B†	160	4		4 th	IM	IM				2.2.09
2220		unknown	170			1 th	C	U				7.0.0
2230	*trmU*	tRNA (5-methylaminomethyl-2-thiouridylate)-methyltransferase	370				E	U			2.1.1.61	2.2.01
2240		membrane protein*	369	1‡		1 th	C	U				4.1.1
2250		membrane protein*	347	1‡		1 th	IM	U				4.1.1
2260		membrane protein*	313	1‡		1 th	E	U				4.1.1
2270		membrane protein*	384	1‡		1 th	E	U				4.1.1
2280		membrane protein*	341	1‡		1 th	C	U				4.1.1
2290		membrane protein*	342	1‡		1 th	E	U				4.1.1
2300		membrane protein*	370	1		1 th	C	U	+			4.1.1
2310		exported protein*	317		signal	1 th	C	C				4.1.0
2320		exported protein*	307		signal	1 th	C	U				4.1.0
2330		membrane protein*	306	1‡		1 th	IM	U				4.1.1
2340		membrane protein*	326	1‡		1 th	C	U				4.1.1
2370		unknown	417				E	U				0.0.0
2380		unknown	332			signal	OM	U				0.0.0
2390	*uvrD*	DNA helicase II	639				C	U			3.6.1.-	2.2.03
2400		membrane protein*	391	1‡		1 th	E	U		90 bp x 2.0		4.1.1
2410		membrane protein*	326	1‡		1 th	C	U				4.1.1
2420	*gyrA*	DNA gyrase subunit A	898				OM	C			5.99.1.3	2.2.03
2430	*nth*	endonuclease III	210				C	U			4.2.99.18	2.2.03

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
2440		integral membrane protein*	220	2		2 th	IM	U				4.1.1
2450	*htpG*	chaperone protein HtpG	637				C	C				1.3.1
2460	*purB*	adenylosuccinate lyase	432				C	C			4.3.2.2	3.7.1
2470		integral membrane protein*	358	2		4 th	C	IM				4.1.1
2480		integral membrane protein*	368	3		4 th	IM	IM				4.1.1
2490		unknown	831			3 th	OM	IM				0.0.0
2500		unknown	305				OM	U				0.0.0
2510		unknown	807			2 th	OM	OM				0.0.0
2520		biotin--[acetyl-CoA-carboxylase] synthetase*	252				E	U			6.3.4.15	3.2.02
2530		glutathione S-transferase*	241				C	C		155 bp x 3.0 (C-terminus)	2.5.1.18	3.4.5
2540		exported protein*	175		signal	signal	C	U				4.1.0
2550		ABC transporter, ATP-binding protein*	340				C	C				1.5.7
2560	*tatA*	Sec-independent protein translocase membrane protein†	56	1		1 th	C	U				1.5.5
2570	*recR*	recombination protein RecR*	195				C	U				2.2.03
2580		ABC transporter, periplasmic solute binding protein*	287	1‡		signal	C	P				1.5.7
2590		ABC transporter, ATP-binding protein*	242				C	IM				1.5.7
2600	*ubiB*	ubiquinone biosynthesis protein UbiB*	480	1		2 th	IM	C		221 bp x 2.0 (C-terminus)		3.2.08
2610		integral membrane protein*	401	12		12 th	IM	IM		221 bp x 2.0 (N-terminus)		4.1.1
2620		conserved hypothetical protein	445				C	C				0.0.2
2630		unknown	1202				OM	OM		375 bp x 2.1 (C-terminus)		0.0.0
2640		conserved hypothetical protein	274				E	U				0.0.1
2650	*sucA*	2-oxoglutarate dehydrogenase E1 component	913				OM	C			1.2.4.2	3.5.8
2660		unknown	411				E	U				5.1.2
2670	*dapA*	dihydrodipicolinate synthase	296				OM	U			4.2.1.52	3.1.13
2680		HIT-like protein*	113				C	C				7.0.0
2690		unknown	352				E	U				7.0.0
2700	*mutS*	DNA mismatch repair protein MutS	804				OM	C				2.2.03
2710	*nadE*	glutamine-dependent NAD(+) synthetase*	513				OM	U			6.3.5.1	3.2.11
2720	*hemB*	delta-aminolevulinic acid dehydratase	329				C	U			4.2.1.24	3.2.06
2730		unknown	912				OM	C				0.0.0
2740		integral membrane transport protein*	426	12		12 th	IM	IM				1.5.5
2750		membrane protein*	527	1	signal	2 th	E	IM				4.1.1
2760		membrane protein*	519	1	signal	2 th	OM	U				4.1.1
2770		membrane protein*	526	1	signal	2 th	C	U				4.1.1
2780		membrane protein*	524	1	signal	2 th	C	U		21 bp x 2.0		4.1.1
2790		integral membrane protein*	653	2‡		2 th	C	OM				4.1.1

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
2800		membrane protein*	520	1	signal	2 th	C	U		15 bp x 2.0		4.1.1
2810		integral membrane transport protein*	417	12		12 th	IM	IM				1.5.5
2820		integral membrane transport protein*	415	12		12 th	IM	IM				1.5.5
2830	*ssb*	single-strand DNA binding protein	156				C	C				2.2.03
2840	*matA*	malonyl-CoA decarboxylase*	460				C	C			4.1.1.9	3.6.1
2850	*gatB*	aspartyl/glutamyl-tRNA amidotransferase subunit B	481				C	C			6.3.5.-	2.2.01
2860	*fabI*	enoyl-[acyl-carrier-protein] reductase [NADH]	273				C	U			1.3.1.9	3.6.1
2870	*dnaA*	chromosomal replication initiator protein DnaA	464				C	C				2.2.03
2900		integral membrane protein*	331	2		2 th	C	C				4.1.1
2910	*nadD*	nicotinate-nucleotide adenylyltransferase*	194			signal	OM	U			2.7.7.18	3.2.11
2920	*pdxJ*	pyridoxal phosphate biosynthetic protein PdxJ	238				C	C				3.2.12
2930	*hupB*	DNA-binding protein HU-beta*	94				C	U				2.2.03
2940	*holB*	DNA polymerase III, delta prime subunit†	296				C	C			2.7.7.7	2.2.03
2950		conserved hypothetical protein	199				C	U		144 bp x 6.0 (N-terminus)		0.0.2
2960	*coaE*	dephospho-CoA kinase*	201				C	U			2.7.1.24	3.2.17
2970	*thiC*	thiamine biosynthesis protein ThiC	555				C	U				3.2.14
2980		unknown	186				E	U				7.0.0
2990	*rpoZ*	DNA-directed RNA polymerase omega chain*	132				C	C			2.7.7.6	2.2.11
3000		unknown	123				C	U				7.0.0
3010	*leuS*	leucyl-tRNA synthetase	830				C	C			6.1.1.4	2.2.01
3030		deoxyribonuclease†	261				C	C			3.1.21.-	7.0.0
3040	*pyrF*	orotidine 5'-phosphate decarboxylase	231				C	IM			4.1.1.23	3.7.2
3050	*surE*	acid phosphatase SurE	252				E	U			3.1.3.2	1.4.0
3060	*ccmE*	cytochrome c-type biogenesis protein CcmE	134	1‡		signal	P	U				2.2.13
3070	*nuoC*	NADH-quinone oxidoreductase chain C*	191				C	C			1.6.99.5	3.5.3
3090	*nuoB*	NADH-quinone oxidoreductase chain B	172				P	U			1.6.99.5	3.5.3
3100	*nuoA*	NADH-quinone oxidoreductase chain A*	123	3		3 th	IM	IM			1.6.99.5	3.5.3
3110	*uvrA*	uvrABC system protein A	959				OM	OM				2.2.03
3120		unknown	174				C	U				7.0.0
3130	*ribH*	6,7-dimethyl-8-ribityllumazine synthase*	149				C	U			2.5.1.9	3.2.13
3140		integral membrane protein*	575	6‡		6 th	IM	IM				4.1.1
3150		integral membrane transport protein*	461	11		10 th	IM	IM				1.5.5
3160	*pssA*	CDP-diacylglycerol--serine O-phosphatidyltransferase*	260	7			IM	IM			2.7.8.-	2.2.07
3170	*psd*	phosphatidylserine decarboxylase proenzyme*	227	1		1 th	IM	U			4.1.1.65	2.2.07
3180		unknown	1134				C	U		182 bp x 4.4 (C-terminus)		7.0.0
3190	*efp*	elongation factor P*	189				C	C				2.2.10
3200	*suhB*	inositol-1-monophosphatase*	256				OM	U			3.1.3.25	6.1.3

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
3210	*rluC*	ribosomal large subunitPSeudouridine synthase C*	305				C	C			4.2.1.70	2.2.11
3220		response regulator component of a two-component regulatory system†	461				C	C				6.0.0
3221		unknown	93				C	U				0.0.0
3230		NAD-glutamate dehydrogenase†	1589				C	OM			1.4.1.-	3.3.20
3240		integral membrane protein*	210	4		4 th	IM	IM				4.1.1
3250	*cysS*	cysteinyl-tRNA synthetase	457				C	C			6.1.1.16	2.2.01
3260		unknown	648				OM	U				0.0.0
3270	*nrdB*	ribonucleoside-diphosphate reductase beta chain*	324	1		1 th	C	U			1.17.4.1	3.7.0
3280		conserved hypothetical protein	94				P	U				0.0.1
3290		unknown	194				E	U				0.0.0
3300		conserved hypothetical protein	339				C	U				0.0.2
3310	*dnaG*	DNA primase*	592				C	C			2.7.7.-	2.2.03
3320	*rpoD*	RNA polymerase sigma-70 factor	622				C	C	+++			2.2.11
3330		conserved hypothetical protein	317				C	U				0.0.2
3340	*ispE*	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase*	281				C	U			2.7.1.148	3.2.08
3350	*cutA*	periplasmic divalent cation tolerance protein CutA*	109				C	C				1.6.4
3360		two component sensor kinase*	828	3		3 th	C	IM				6.1.0
3370	*mdmC*	O-methyltransferase*	218				E	U			2.1.1.-	2.2.0
3380		unknown	94				C	U				0.0.0
3390		conserved hypothetical protein	467			signal	IM	U				0.0.2
3400	*topA*	DNA topoisomerase I	819				OM	U			5.99.1.2	2.2.03
3410		unknown	119				C	U				0.0.0
3420		conserved hypothetical protein	153				C	U				0.0.1
3430	*acpS*	holo-[acyl-carrier-protein] synthase*	123				C	U			2.7.8.7	3.6.1
3440	*proS*	prolyl-tRNA synthetase	426				C	C			6.1.1.15	2.2.01
3450		exported protein*	327		signal	signal	C	U				4.1.0
3460	*coaD*	phosphopantetheine adenylyltransferase*	165				C	U			2.7.7.3	3.2.17
3470	*trxB*	thioredoxin reductase	318			signal	C	U			1.8.1.9	3.2.15
3480		peroxiredoxin*	205				C	C				1.4.1
3490	*aatA*	aspartate aminotransferase A	399				OM	U			2.6.1.1	3.1.07
3500	*ppiD*	peptidyl-prolyl cis-trans isomerase D*	630	1‡		signal	OM	OM	+		5.2.1.8	1.3.1
3510		glycoprotease†	193				C	U			3.4.-.-	2.1.4
3520	*truB*	tRNAPSeudouridine synthase B*	296				C	C			4.2.1.70	2.2.11
3530	*rpsO*	30S ribosomal protein S15	93				C	U				4.2.2
3540	*pnp*	polyribonucleotide nucleotidyltransferase	789				OM	C			2.7.7.8	2.1.2
3550		conserved hypothetical protein	265				IM	IM				0.0.2
3560	*lepA*	GTP-binding protein LepA	598				C	C				2.2.10
3570		integral membrane protein*	376	2		2 th	E	U		12 bp x 3.2		4.1.1

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
3580		integral membrane protein*	188	2‡		signal, 1 th	OM	U				4.1.1
3590		integral membrane protein*	389	2		2 th	E	U		45 bp x 7.4, 42 bp x 2.1		4.1.1
3600		integral membrane protein*	585	2‡		2 th	OM	IM		12 bp x 16.7		4.1.1
3610		membrane protein*	513	1	signal	2 th	C	U				4.1.1
3620		integral membrane protein*	537	2‡		2 th	OM	U				4.1.1
3630		membrane protein*	519	1	signal	signal, 1 th	OM	U				4.1.1
3640		unknown	111				C	U				0.0.0
3650	*prfB*	peptide chain release factor 2	367				OM	C				2.2.10
3660		conserved hypothetical protein	249				C	U				0.0.2
3670	*gatA*	glutamyl-tRNA(Gln) amidotransferase subunit A	487				OM	U			6.3.5.7	2.2.01
3680	*folC*	folylpolyglutamate synthase/dihydrofolate synthase*	431				C	U			6.3.2.17, 6.3.2.12	3.2.05
3690	*hemC*	porphobilinogen deaminase	299				C	U			2.5.1.61	3.2.06
3700	*typA*	GTP-binding protein TypA/BipA*	612				C	C				7.0.0
3701		unknown	106				C	U				0.0.0
3710	*nuoI*	NADH-quinone oxidoreductase chain I	168	1		1 th	C	C			1.6.99.5	3.5.3
3720	*sipF*	prokaryotic type I signal peptidase	238	1		signal	C	IM			3.4.21.89	2.2.10
3730		unknown	153				C	U		27 bp x 2.2		7.0.0
3740		metal dependent phosphohydrolase*	403				C	C				7.0.0
3750		unknown	1674				OM	C		27 bp x 8.3, 144 bp x 3.9		7.0.0
3760		integral membrane protein*	308	6		6 th	IM	IM				4.1.1
3770	*argG*	argininosuccinate synthase	394				C	C			6.3.4.5	3.1.02
3780		exported protein*	223		signal	signal	OM	IM				4.1.0
3790		exported protein*	235		signal	signal	OM	OM				4.1.0
3800	*argJ*	arginine biosynthesis bifunctional protein ArgJ	419				E	C			2.3.1.1, 2.3.1.35	3.1.02
3810	*exoA*	exodeoxyribonuclease*	278				C	C			3.1.11.2	2.2.03
3820		integral membrane protein*	260	6		6 th	IM	IM				4.1.1
3830		integral membrane protein*	276	6		6 th	IM	IM				4.1.1
3840	*fabG*	3-oxoacyl-[acyl carrier protein] reductase	245				C	C			1.1.1.100	3.6.1
3850	*putA*	proline dehydrogenase/delta-1-pyrroline-5-carboxylate dehydrogenase	1043			1 th	C	C			1.5.99.8, 1.5.1.12	3.1.16
3860		membrane protein*	171	1‡		1 th	E	U				4.1.1
3870	*bioA*	adenosylmethionine-8-amino-7-oxononanoate aminotransferase	425				C	U			2.6.1.62	3.2.02
3880		conserved hypothetical protein	471				OM	U				0.0.2
3890		unknown	126				P	U				0.0.0
3900		unknown	189				C	U				0.0.0
3910		unknown	129				C	U				0.0.0

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
3920		unknown	136				C	U				0.0.0
3930		unknown	188				C	C				0.0.0
3940		unknown	115				C	C				0.0.0
3950	*rpmJ*	50S ribosomal protein L36	42				C	U				4.2.2
3960	*rpoH*	RNA polymerase sigma-32 factor	296				C	C	+			2.2.11
3970		unknown	198				C	C				7.0.0.
3980		unknown	3002				OM	E		144 bp x 2.7, 36 bp x 2.7, 93 bp x 9.2		7.0.0
3990	*atpG*	ATP synthase gamma chain	283			1 th	OM	U			3.6.3.14	3.5.9
4000	*folE*	GTP cyclohydrolase I	190				C	IM			3.5.4.16	3.2.05
4010	*pmbA*	PmbA protein*	455				OM	C				1.3.1
4020		pyridine nucleotide-disulphide oxidoreductase*	337	1			OM	C				3.3.00
4030	*ksgA*	dimethyladenosine transferase	262	1			C	U			2.1.1.-	2.2.11
4040	*tpiA*	triosephosphate isomerase	240				OM	U			5.3.1.1	3.3.04
4050		exported protein*	326		signal	signal	C	IM				4.1.0
4060	*gcp*	o-sialoglycoprotein endopeptidase	348				C	E			3.4.24.57	2.1.4
4070		integral membrane protein*	193	3		3 th	IM	IM				4.1.1
4080	*folB*	dihydroneopterin aldolase†	115				C	U			4.1.2.25	3.2.05
4090	*mdh*	malate dehydrogenase	314				C	C			1.1.1.37	3.5.8
4100	*rpiB*	ribose 5-phosphate isomerase B	146				C	C			5.3.1.6	3.3.09
4110	*ubiG*	3-demethylubiquinone-9 3-methyltransferase*	241				C	C			2.1.1.64	3.2.08
4120		conserved hypothetical protein	156				C	U				0.0.2
4130		conserved hypothetical protein	240				P	U				0.0.2
4140		unknown	522				OM	U				7.0.0
4150	*iscS*	cysteine desulfurase	413				C	C			4.4.1.-	2.2.11
4160		NifU-like protein*	137				P	U				3.3.19
4170		conserved hypothetical protein	149				C	U				0.0.2
4180	*hscB*	co-chaperone protein HscB†	145				C	U				1.3.1
4190	*hscA*	chaperone protein HscA	616				C	C				1.3.1
4200	*fdxB*	ferredoxin, 2FE-2S	122				C	C				3.5.3
4210		membrane protein*	356	1‡		signal	OM	U				4.1.1
4211		cytochrome c-type biogenesis protein†	125	1		1 th	C	U				2.2.13
4220	*lysS*	lysyl-tRNA synthetase	512				C	C		21 bp x 2.0	6.1.1.6	2.2.01
4230		integral membrane protein*	135	2		signal	C	U				4.1.1
4240	*truA*	tRNAPSeudouridine synthase A	246				C	U			4.2.1.70	2.2.01
4250	*pyrB*	aspartate carbamoyltransferase	306				C	C			2.1.3.2	3.7.2
4260	*gyrB*	DNA gyrase subunit B	798				C	U			5.99.1.3	2.2.03
4261		unknown	84				P	U				0.0.0

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
4270	*nuoG*	NADH-quinone oxidoreductase chain G	684				C	C			1.6.99.5	3.5.3
4280	*nuoH*	NADH-quinone oxidoreductase chain H	367	8		8 th	IM	IM			1.6.99.5	3.5.3
4310	*gltX2*	glutamyl-tRNA synthetase 2	470				C	C			6.1.1.17	2.2.01
4320		unknown	425				C	C				0.0.0
4330	*mutM*	formamidopyrimidine-DNA glycosylase	269				C	U			3.2.2.23	2.2.03
4340		unknown	392				OM	U				0.0.0
4350		unknown	409				OM	U				0.0.0
4360		unknown	157				OM	U				0.0.0
4370	*miaA*	tRNA delta(2)-isopentenylpyrophoshate transferase*	300				C	U			2.5.1.8	2.2.11
4390		unknown	240				C	C				0.0.0
4400		unknown	994				C	U				0.0.0
4410		type IV secretion system protein†	232	1		1 th	OM	OM				1.5.6
4420	*nuoD*	NADH-quinone oxidoreductase chain D	393				C	C			1.6.99.5	3.5.3
4430	*nuoE*	NADH-quinone oxidoreductase chain E	183				C	C			1.6.99.5	3.5.3
4440		integral membrane protein*	195	#4		4 th	C	IM				4.1.1
4450		unknown	280				OM	U				0.0.0
4460	*pccB*	propionyl-CoA carboxylase beta chain	510				C	U			6.4.1.3	3.4.4
4470		exported protein*	385		signal	signal	OM	OM				4.1.0
4480	*argB*	acetylglutamate kinase	305				C	U			2.7.2.8	3.1.02
4490	*engB*	GTP binding protein EngB*	200				C	U				1.7.1
4500	*prfA*	peptide release factor 1	359				C	C				2.2.10
4510		sodium:dicarboxylate symporter (glutamate)*	402	8		8 th	IM	IM				1.5.1
4520	*rmuC*	DNA recombination protein RmuC	436	1		signal	C	U				2.2.03
4530		unknown	199				E	U		22 bp x 2.0		0.0.0
4540	*serS*	seryl-tRNA synthetase	427				C	C			6.1.1.11	2.2.01
4550	*hemF*	coproporphyrinogen III oxidase	288				E	U			1.3.3.3	3.2.06
4560		conserved hypothetical protein	95				C	C				0.0.1
4570	*tal*	transaldolase*	220				C	U			2.2.1.2	3.3.09
4580	*atpC*	ATP synthase epsilon chain†	134				C	C			3.6.3.14	3.5.9
4590	*atpD*	ATP synthase beta chain	504				C	U			3.6.3.14	3.5.9
4600		magnesium transporter*	456	4		5 th	IM	IM				1.5.2
4610		membrane protein*	124	1‡		1 th	C	C				4.1.1
4620		membrane protein*	134	1‡		1 th	IM	U				4.1.1
4630		membrane protein*	125	1‡		1 th	P	U				4.1.1
4640		membrane protein*	123	1		1 th	P	U				4.1.1
4650		unknown	771				C	C				0.0.0
4660	*clpA*	ATP-dependent Clp protease, ATP-binding subunit	764				C	C			3.4.21.92	2.1.4
4670		conserved hypothetical integral membrane protein	235	7		7 th	IM	IM				4.1.1
4680	*rbfA*	ribosome-binding factor A*	115				C	C				2.2.11

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
4690	*infB*	translation initiation factor IF-2	856				C	C				2.2.10
4700	*nusA*	N utilization substance protein A*	517				C	C				2.2.11
4710		integral membrane transport protein*	1039	12		12 th	IM	IM				1.5.5
4720	*tatC*	Sec-independent protein translocase protein TatC	250	5		6 th	IM	IM				1.5.5
4730	*ispG*	1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase*	422				C	U		247 bp x 2.1 (N-terminus)	1.17.4.3	3.2.08
4740		exported protein*	639		signal	signal	E	OM		138 bp x 6.9		4.1.0
4750	*dxr*	1-deoxy-D-xylulose 5-phosphate reductoisomerase	387	1		1 th	IM	U			1.1.1.267	3.2.08
4760	*nuoN*	NADH-quinone oxidoreductase chain N	474	13		14 th	IM	IM			1.6.99.5	3.5.3
4770	*nuoM*	NADH-quinone oxidoreductase chain M	486	14		14 th, signal	IM	IM			1.6.99.5	3.5.3
4780	*nuoL*	NADH-quinone oxidoreductase chain L	622	16		17 th, signal	IM	IM	+		1.6.99.5	3.5.3
4790	*nuoK*	NADH-quinone oxidoreductase chain K	108	3		3 th	IM	IM			1.6.99.5	3.5.3
4800	*nuoJ*	NADH-quinone oxidoreductase chain J	200	5		5 th	IM	IM			1.6.99.5	3.5.3
4810	*nuoF*	NADH-quinone oxidoreductase chain F	425			1 th	C	C			1.6.99.5	3.5.3
4820	*rpmA*	50S ribosomal protein L27	88				C	U				4.2.2
4830	*rplU*	50S ribosomal protein L21	102				C	U				4.2.2
4840	*eno*	enolase	421				C	C			4.2.1.11	3.3.15
4850		conserved hypothetical GTP-binding protein	340				C	C		9 bp x 3.0		0.0.2
4860	*mraW*	S-adenosyl-methyltransferase MraW*	301				OM	C			2.1.1.-	7.0.0
4870	*ileS*	isoleucyl-tRNA synthetase	1104				C	C			6.1.1.5	2.2.01
4880		bacterioferritin comigratory protein†	147				C	U				7.0.0
4890		conserved hypothetical protein	238				C	C				0.0.2
4900		unknown	173				C	C				7.0.0
4910	*argS*	arginyl-tRNA synthetase	576			1 th	C	C			6.1.1.19	2.2.01
4920	*recO*	DNA repair protein RecO†	244				C	U				2.2.03
4930		unknown	186				E	U				0.0.0
4950		competence protein†	492				C	C				7.0.0
4960		integral membrane protein*	129	2		2 th	P	U				4.1.1
4970	*rbn*	tRNA processing ribonuclease BN†	277	5		6 th	IM	IM				2.2.12
4980	*thiL*	thiamine-monophosphate kinase*	316				C	IM			2.7.4.16	3.2.14
4990	*dnaQ*	DNA polymerase III, epsilon subunit	242				C	C			2.7.7.7	2.2.03
5000		exported protein*	490		signal	signal	OM	OM				4.1.0
5010		exported protein*	564		signal	signal	OM	OM		24 bp x 2.1		4.1.0
5020	*petC*	cytochrome c1 precursor	252	1	signal	1 th, signal	P	U				3.5.3
5030	*petB*	cytochrome b	408	9		9 th	IM	IM		20 bp x 2.0		3.5.3
5040	*petA*	ubiquinol-cytochrome c reductase iron-sulphur subunit	187	1		1 th	C	U			1.10.2.2	3.5.3
5050		integral membrane protein*	290	5		5 th	IM	IM				4.1.1
5060		ABC transporter, membrane-spanning protein*	266	7		8 th	IM	IM				1.5.7

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
5070		conserved hypothetical protein	120				C	C				0.0.2
5080	*tsf*	elongation factor Ts	288				C	C				2.2.10
5090	*rpsB*	30S ribosomal protein S2	286				C	U				4.2.2
5100	*maf*	septum formation protein Maf†	192				OM	U				7.0.0
5110	*infA*	translation initiation factor IF-1*	82				C	C				2.2.10
5120		secretion protein†	363	1		1 th	OM	U				7.0.0
5130		dihydrolipoamide dehydrogenase, E3 component of pyruvate or 2-oxoglutarate dehydrogenase complex*	465				C	C			1.8.1.4	3.5.0
5140		exported protein*	389		signal	signal	OM	U				4.1.0
5150	*gpmI*	2,3-bisphosphoglycerate-independent phosphoglycerate mutase	501				C	C			5.4.2.1	3.3.15
5160	*engA*	GTP binding protein EngA*	439				C	U				7.0.0
5170	*carA*	carbamoyl-phosphate synthase small chain	374				E	U			6.3.5.5	3.1.02
5180	*ispH*	4-hydroxy-3-methylbut-2-enyl diphosphate reductase	328				C	U			1.17.1.2	3.2.08
5190	*dut*	deoxyuridine 5'-triphosphate nucleotidohydrolase*	124				C	U			3.6.1.23	3.3.11
5200		conserved hypothetical protein	203				C	U				0.0.1
5210		type IV secretion system protein†	2455	4	signal	6 th, signal	OM	OM		261 bp x 5.1, 222 bp x 3.9, 180 bp x 1.9		1.5.6
5220		type IV secretion system protein†	1529	6	signal	5 th, signal	OM	OM		216 bp x 2.4, 15 bp x 2.3		1.5.6
5230		type IV secretion system protein†	911	7	signal	7 th, signal	OM	IM				1.5.6
5240	*virB6*	type IV secretion system protein VirB6	821	8	signal	9 th, signal	IM	IM				1.5.6
5250	*virB4*	type IV secretion system protein VirB4	800				C	U				1.5.6
5260	*virB3*	type IV secretion system protein VirB3	97	2		2 th	IM	U				1.5.6
5270	*sodB*	superoxide dismutase [Fe]	210				E	U			1.15.1.1	1.4.2
5280		ABC transporter, membrane-spanning protein*	420	6		6 th	IM	IM				1.5.7
5290	*lipA*	lipoic acid synthetase	292				C	C				3.2.07
5300		unknown	464				OM	U				0.0.0
5310		integral membrane protein*	1392	2		3 th	C	C				4.1.1
5320	*bccA*	acetyl-/propionyl-coenzyme A carboxylase alpha chain*	660				C	C		14 bp x 2.1	6.3.4.14	3.6.1
5330	*rluD*	ribosomal large subunitPSeudouridine synthase D	324				C	U			4.2.1.70	2.2.11
5340	*lysA*	diaminopimelate decarboxylase*	420				C	U			4.1.1.20	3.1.13
5350	*rpmB*	50S ribosomal protein L28	100				C	U				4.2.2
5360	*priA*	primosomal protein N'	659				OM	U				2.2.03
5370		exported protein*	325		signal	signal	C	U				4.1.0
5380	*hemD*	uroporphyrinogen-III synthase*	242				C	U			4.2.1.75	3.2.06
5390		aminopeptidase†	572				C	U				7.0.0
5400		unknown	173			1 th	OM	U				7.0.0
5410		conserved hypothetical protein	275				C	C				0.0.2
5420	*era*	GTP-binding protein ERA*	296				C	U				7.0.0

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
5430	*ffh*	signal recognition particle protein	450				C	C				1.5.5
5440		NADH-quinone oxidoreductase subunit*	528	15		16 th	IM	IM			1.6.99.5	3.5.3
5450		unknown	264				OM	U				0.0.0
5460		unknown	258				C	U				0.0.0
5470		membrane protein*	158	1		1 th	C	U				4.1.1
5480		membrane protein*	111	1‡		signal	C	IM				4.1.1
5490		conserved hypothetical protein	153				C	U				0.0.2
5500	*dnaK*	chaperone protein DnaK	645				C	U				1.3.1
5510		ribonuclease*	610				OM	U			3.1.4.-	2.1.2
5520		integral membrane protein*	111	3		3 th	IM	IM				4.1.1
5530		Na+/H+ antiporter subunit*	139	4		4 th	IM	IM				1.5.2
5540		integral membrane protein*	181	6		6 th	IM	IM				4.1.1
5550		Na+/H+ antiporter subunit*	99	3		3 th	IM	IM				1.5.2
5560		integral membrane protein*	88	3		2 th, signal	IM	IM				4.1.1
5570		unknown	552				OM	U		183 bp x 3.0		0.0.0
5580		unknown	344				E	U				0.0.0
5590		unknown	213				C	C				0.0.0
5600	*tkt*	transketolase	671				C	U			2.2.1.1	3.3.09
5610		carboxypeptidase†	491				C	U				2.1.4
5620		unknown	217			signal	P	OM				7.0.0
5630	*purA*	adenylosuccinate synthetase	430				C	C			6.3.4.4	3.7.1
5640		Holliday junction resolvase†	156				C	U			3.1.-.-	2.2.03
5650	*nrdA*	ribonucleoside-diphosphate reductase alpha chain*	595				C	U			1.17.4.1	3.7.0
5660	*ispA*	geranyltranstransferase*	276				C	C			2.5.1.10	3.2.08
5670		membrane protein*	317	1‡		signal	C	U				4.1.1
5680	*thiO*	thiamine biosynthesis oxidoreductase*	354		signal		C	C				3.2.14
5690		deaminase†	145				C	U			3.5.4.-	7.0.0
5700		membrane protein*	142	1		1 th	C	U				4.1.1
5710	*dnaB*	replicative DNA helicase	486				C	C	+		3.6.1.-	2.2.03
5720	*fabH*	3-oxoacyl-[acyl-carrier-protein] synthase III	319	1		1 th	IM	U			2.3.1.41	3.6.1
5730	*plsX*	fatty acid/phospholipid synthesis protein	336				C	C				2.2.07
5740	*rpmF*	50S ribosomal protein L32	60				C	U				4.2.2
5750	*tgt*	queuine tRNA-ribosyltransferase	378				C	U			2.4.2.29	2.2.12
5760	*pstB*	phosphate ABC transporter, ATP-binding protein*	253				C	U			3.6.3.27	1.5.7
5770	*dapB*	dihydrodipicolinate reductase	264				C	U			1.3.1.26	3.1.13
5780		monooxygenase*	164				C	U				7.0.0
5790	*ubiA*	4-hydroxybenzoate octaprenyltransferase	295	8		8 th	IM	IM			2.5.1.-	3.2.08
5791	*rpmH*	50S ribosomal protein L34	44				C	U				4.2.2
5800	*rnpA*	ribonuclease P protein component*	122				C	U			3.1.26.5	2.2.11

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
5810		integral membrane transport protein*	437	11		11 th	IM	IM				1.5.5
5820		competence protein†	650	12		11 th	IM	IM		142 bp x 10 (C-terminus)		1.5.5
5830	*pheT*	phenylalanyl-tRNA synthetase beta chain	789				C	C			6.1.1.20	2.2.01
5840	*rplQ*	50S ribosomal protein L17	128				C	C				4.2.2
5850	*rpoA*	DNA-directed RNA polymerase alpha chain	358				C	C			2.7.7.6	2.2.11
5860	*rpsK*	30S ribosomal protein S11	127				P	U				4.2.2
5870	*rpsM*	30S ribosomal protein S13	123				C	C				4.2.2
5880	*adk*	adenylate kinase	220				C	C			2.7.4.3	3.7.1
5890	*secY*	preprotein translocase secY subunit	432	11		11 th	IM	IM				1.5.5
5900	*rplO*	50S ribosomal protein L15	157				C	U				4.2.2
5910	*rpsE*	30S ribosomal protein S5	174				C	C				4.2.2
5920	*rplR*	50S ribosomal protein L18	120				C	C				4.2.2
5930	*rplF*	50S ribosomal protein L6	178				OM	U				4.2.2
5940	*rpsH*	30S ribosomal protein S8	132				C	C				4.2.2
5950	*rpsN*	30S ribosomal protein S14	101				C	U				4.2.2
5960	*rplE*	50S ribosomal protein L5	177				C	U				4.2.2
5970	*rplX*	50S ribosomal protein L24	109				P	U				4.2.2
5980	*rplN*	50S ribosomal protein L14	119				C	U				4.2.2
5990	*rpsQ*	30S ribosomal protein S17	74				C	U				4.2.2
5991	*rpmC*	50S ribosomal protein L29	67				C	C				4.2.2
6000	*rplP*	50S ribosomal protein L16	136				P	U				4.2.2
6010	*rpsC*	30S ribosomal protein S3	211				C	U				4.2.2
6020	*rplV*	50S ribosomal protein L22	114				C	U				4.2.2
6030	*rpsS*	30S ribosomal protein S19	93				P	U				4.2.2
6040	*rplB*	50S ribosomal protein L2	276				E	U				4.2.2
6050	*rplW*	50S ribosomal protein L23	96				C	U				4.2.2
6060	*rplD*	50S ribosomal protein L4	205				C	U				4.2.2
6070	*rplC*	50S ribosomal protein L3	231				P	U				4.2.2
6080	*rpsJ*	30S ribosomal protein S10	111				C	U				4.2.2
6090	*tufB*	elongation factor Tu-B	395				C	C				2.2.10
6100		tRNA/rRNA methyltransferase*	249				C	U			2.1.1.-	2.2.11
6110	*cmk*	cytidylate kinase*	212				C	U			2.7.4.14	3.7.2
6120	*rpsA*	30S ribosomal protein S1	565				OM	C				4.2.2
6130		peptidase*	289	1		1 th	OM	U			3.4.21.-	2.1.4
6140	*ihfB*	integration host factor beta subunit†	87				C	U				2.2.03
6150		unknown	97				C	C				0.0.0
6160		unknown	121				C	C				0.0.0
6170		integral membrane protein*	325	5		4 th	IM	IM				4.1.1

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
6180	*hemH*	ferrochelatase	313			signal	C	C			4.99.1.1	3.2.06
6190		ATPase*	357			1 th	C	U				7.0.0
6200		unknown	101				C	U				0.0.0
6210		integral membrane protein*	285	3		3 th	IM	IM				4.1.1
6220		unknown	125				C	U				7.0.0
6230		exported protein*	134		signal	signal	P	U				4.1.0
6240		membrane protein*	81	1‡		signal	P	U				4.1.1
6250	*recB*	exodeoxyribonuclease V beta chain*	857				C	U	+	132 bp x 5.4 (C-terminus)	3.1.11.5	2.2.03
6260	*qor*	quinone oxidoreductase*	324				C	U			1.6.5.5	3.5.3
6270		ABC transporter, ATP-binding protein*	593	5		5 th	IM	IM				1.5.7
6280	*folP1*	dihydropteroate synthase 1*	274				C	U			2.5.1.15	3.2.05
6290	*folP2*	dihydropteroate synthase 2*	283				C	C			2.5.1.15	3.2.05
6300		integral membrane protein*	352	2		2 th	C	U				4.1.1
6310	*carB*	carbamoyl-phosphate synthase, large subunit	1075				C	U			6.3.5.5	3.1.02
6320		unknown	105				C	U				0.0.0
6330	*fumC*	fumarate hydratase class II	461				C	C			4.2.1.2	3.5.8
6350	*pyrC*	dihydroorotase	449				C	U			3.5.2.3	3.7.2
6360	*lipB*	lipoate-protein ligase B	208				C	C			6.-.-.-	2.2.0
6370	*purN*	phosphoribosylglycinamide formyltransferase	212				IM	U			2.1.2.2	3.7.1
6380	*pepA*	cytosol aminopeptidase	500			1 th	OM	U			3.4.11.1	2.1.4
6390		dioxygenase*	244	1		1 th	E	U			1.3.11.-	7.0.0
6400	*clpB*	heat shock protein ClpB	859				C	C				1.3.1
6410		oxidoreductase*	249				OM	C			1.-.-.-	7.0.0
6420	*groEL*	60 kDa chaperonin GroEL	551				C	U				1.3.1
6430	*groES*	10 kDa chaperonin GroES	94				C	C				1.3.1
6440	*radC*	DNA repair protein RadC	230				C	U				2.2.03
6450	*purQ*	phosphoribosylformylglycinamidine synthase I†	265				C	U			6.3.5.3	3.7.1
6460	*gidA*	glucose inhibited division protein A	625				OM	U				1.7.1
6470	*glpX*	fructose-1,6-bisphosphatase class II GlpX	306				C	C			3.1.3.11	3.3.04
6480		peptidase*	204			signal	C	U				7.0.0
6490		unknown	151				C	C				0.0.0
6500	*bioB*	biotin synthase	322				C	C			2.8.1.6	3.2.02
6510	*purL*	phosphoribosylformylglycinamidine synthase II*	1010				C	U		185 bp x 6.2 (C-terminus)	6.3.5.3	3.7.1
6520	*folK*	2-amino-4-hydroxy-6-hydroxymethyldihydropteridine pyrophosphokinase*	174				C	U		124 bp x 4.4 (C-terminus)	2.7.6.3	3.2.05
6530		unknown	210			signal	E	U				0.0.0
6540		zinc metallopeptidase*	433	2‡		signal	IM	IM				7.0.0
6550		conserved hypothetical protein	121				C	U				0.0.2

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
6560		unknown	295				C	U				0.0.0
6570		unknown	212				C	U				0.0.0
6580	*purM*	phosphoribosylformylglycinamidine cyclo-ligase	342				C	U			6.3.3.1	3.7.1
6590		integrase/recombinase XerD or XerC*	312				C	U				2.2.03
6600	*gpsA*	glycerol-3-phosphate dehydrogenase [NAD(P)+]	327			signal	C	IM			1.1.1.94	2.2.07
6610		response regulator component of a two-component regulatory system*	471				C	C	+			6.1.2
6620	*ftsQ*	cell division protein FtsQ*	271	1		1 th	C	C				1.7.1
6640	*gshB*	glutathione synthetase	312				C	C			6.3.2.3	3.2.18
6650		exported protein*	200		signal	signal	IM	P				4.1.0
6660	*aspS*	aspartyl-tRNA synthetase	590				C	C			6.1.1.12	2.2.01
6670		haloacid dehalogenase-like hydrolase*	210				E	U			3.1.3.-	7.0.0
6680		integral membrane protein*	170	3		3 th	IM	IM				4.1.1
6690	*ppdK*	pyruvate phosphate dikinase	873				C	C			2.7.9.1	3.3.15
6700		NADH-quinone oxidoreductase subunit*	492	13		14 th	IM	IM			1.6.99.5	3.5.3
6710		conserved hypothetical protein	258			1 th	C	U				0.0.2
6720		c-type cytochrome*	174	1		1 th	P	IM				3.5.3
6730	*folD*	methylenetetrahydrofolate dehydrogenase/methenyltetrahydrofolate cyclohydrolase	300				C	U			1.5.1.5 3.5.4.9	3.2.05
6740	*gmk*	guanylate kinase	209				C	C			2.7.4.8	3.7.1
6750	*ccmC*	heme exporter protein C	234	6		6 th	IM	IM				1.5.7
6760	*ruvA*	Holliday junction DNA helicase RuvA*	191				C	U				2.2.03
6770	*ruvB*	Holliday junction DNA helicase RuvB	331				C	C				2.2.03
6780	*bcr*	bicyclomycin resistance protein*	398	12		12 th	IM	IM				1.5.5
6790	*thyX*	thymidylate synthase complementing protein*	285				C	U			2.1.1.148	7.0.0
6800	*sdhB*	succinate dehydrogenase iron-sulfur subunit	258				C	U			1.3.99.1	3.5.8
6810	*sdhA*	succinate dehydrogenase flavoprotein subunit	598				C	U			1.3.99.1	3.5.8
6820		ABC transporter, ATP-binding and membrane-spanning protein*	583	5		6 th	OM	IM				1.5.7
6830		unknown	109				C	U				0.0.0
6840	*glyA*	serine hydroxymethyltransferase	421				C	C			2.1.2.1	3.1.09
6850	*rplI*	50S ribosomal protein L9	207				C	C				4.2.2
6860	*rpsR*	30S ribosomal protein S18	95				C	U				4.2.2
6870	*rpsF*	30S ribosomal protein S6	109				C	U				4.2.2
6880		integral membrane protein*	203	3		3 th	IM	IM				4.1.1
6890		integral membrane protein*	881	20		22 th	IM	IM				4.1.1
6900	*radA*	DNA repair protein RadA	450				OM	U				2.2.03
6910	*dsbE*	thiol:disulfide interchange protein*	166	1‡		signal	C	P				2.2.13
6920		conserved hypothetical protein	73				C	C				0.0.2
6930		glutaredoxin-related protein†	110				C	C				7.0.0

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
6940	*ligA*	NAD-dependent DNA ligase	674				OM	U		238 bp x 3.2 (C-terminus)	6.5.1.2	2.2.03
6950		exonuclease*	389				OM	U				7.0.0
6960		histidine kinase sensor component of a two-component regulatory system*	710	5		5 th	IM	IM				6.1.2
6970		unknown	95				C	C				7.0.0
6980		conserved hypothetical protein	280				E	U				0.0.2
6990	*dcd*	deoxycytidine triphosphate deaminase*	185				P	U			3.5.4.13	3.7.2
7000	*purC*	phosphoribosylaminoimidazole-succinocarboxamide synthase	250				C	U			6.3.2.6	3.7.1
7010	*hisS*	histidyl-tRNA synthetase	414				OM	C			6.1.1.21	2.2.01
7020		integral membrane protein*	987	2		2 th	IM	C				4.1.1
7030		disulfide oxidoreductase*	250	1‡		signal	C	U				1.3.1
7040		cytochrome c oxidase assembly protein*	359	8		8 th	IM	IM				2.2.13
7050	*ccmA*	heme exporter protein A	213				C	U				1.5.7
7060		unknown	546				C	C				0.0.0
7070		membrane protein*	1373	1		1 th	OM	C		141 bp x 3.9, 198 bp x 5.2		4.1.1
7080		membrane protein*	899	1		1 th	OM	U				4.1.1
7090		membrane protein*	228	1‡		1 th	C	C				4.1.1
7100		membrane protein*	250	1‡		1 th	C	U				4.1.1
7110		exported protein*	182		signal	1 th	C	U				4.1.0
7120		exported protein*	204		signal	1 th	C	C				4.1.0
7130		membrane protein*	186	1‡		1 th	C	C				4.1.1
7140		membrane protein*	197	1‡		1 th	OM	C				4.1.1
7150		membrane protein*	142	1		1 th	C	U				4.1.1
7160		membrane protein*	172	1		1 th	C	C				4.1.1
7170		methylpurine-DNA glycosylase*	188				C	U		178 bp x 2.8 (C-terminus)	3.2.2.-	2.2.03
7180		membrane protein*	241	1‡		1 th	C	U				4.1.1
7190		unknown	281				OM	C				0.0.0
7200		unknown	360				C	C				0.0.0
7220		cytidylyltransferase*	228	6		6 th	IM	IM	+	181 bp x 1.9 (C-terminus)	2.7.7.41	3.6.1
7230	*frr*	ribosome recycling factor	185				C	C				2.2.10
7240	*pyrH*	uridylate kinase	244				C	U			2.7.4.-	3.3.11
7250		membrane protein*	999	1		1 th	OM	OM				4.1.1
7260	*rnhA*	ribonuclease HI	146				C	U			3.1.26.4	2.1.2
7270		membrane protein*	198	1‡		1 th	P	U				4.1.1
7280		membrane protein*	181	1		1 th	C	U				4.1.1
7290	*mfd*	transcription-repair coupling factor	1122				C	U				2.2.03
7300		integral membrane protein*	157	2		signal, 1 th	E	U				4.1.1

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
7310		integral membrane protein*	202	2		signal, 1 th	E	U				4.1.1
7320		integral membrane protein*	266	2		1 th	OM	U				4.1.1
7330		membrane protein*	291	1		signal, 1 th	OM	U				4.1.1
7340		membrane protein*	122	1		1 th	C	U				4.1.1
7350		membrane protein*	145	1		1 th	E	U				4.1.1
7360		membrane protein*	147	1		1 th	IM	U				4.1.1
7370		integral membrane protein*	169	2		2 th	E	U				4.1.1
7380		membrane protein*	157	1		1 th	C	C				4.1.1
7390	*ribE*	riboflavin synthase, alpha subunit*	202				E	U			2.5.1.9	3.2.13
7400		unknown	467				C	U				0.0.0
7410		zinc protease*	421				E	U			3.4.24.-	2.1.4
7420		conserved hypothetical protein	241			signal	C	U	+			0.0.2
7430	*secB*	protein-export protein SecB*	174				C	U				1.3.1
7440		conserved hypothetical protein	213			signal	C	U				0.0.1
7450		integral membrane protein*	409	12		12 th	IM	IM				4.1.1
7460	*tmk*	thymidylate kinase*	202				C	C			2.7.4.9	3.7.2
7470	*fabD*	malonyl CoA-acyl carrier protein transacylase*	320				IM	U			2.3.1.39	3.6.1
7480	*rpmE*	50S ribosomal protein L31	75				C	C				4.2.2
7490	*ppnK*	inorganic polyphosphate/ATP-NAD kinase*	263				OM	U			2.7.1.23	3.3.0
7500	*guaB*	inosine-5'-monophosphate dehydrogenase	485				C	U			1.1.1.205	3.7.1
7510		unknown	281				C	U				7.0.0
7520	*pdhA*	pyruvate dehydrogenase E1 component, alpha subunit	329				C	U			1.2.4.1	3.5.7
7530		conjugal transfer protein*	258			signal	OM	U				1.5.6
7540	*trxA*	thioredoxin 1	107				C	C				3.2.15
7550		conserved hypothetical protein	413				C	C				0.0.2
7560	*xseB*	exodeoxyribonuclease VII small subunit*	62				C	C			3.1.11.6	2.1.1
7570		NADH-ubiquinone oxidoreductase*	97				P	U			1.6.99.3	3.5.3
7580		integral membrane transport protein*	304	10		10 th	IM	IM				1.5.5
7590		conserved hypothetical protein	194				OM	U				0.0.2
7600		membrane protein*	425	1		signal	OM	U		16 bp x 1.9		4.1.1
7610	*gltX1*	glutamyl-tRNA synthetase 1	443				C	C			6.1.1.17	2.2.01
7620		integral membrane protein*	120	3		3 th	IM	U				4.1.1
7630	*thiG*	thiazole biosynthesis protein	261				C	U				3.2.14
7640		thiamin S protein†	74				C	U				3.2.14
7650		unknown	468				E	U				0.0.0
7660		NifU-related protein*	185				C	C				7.0.0
7661		unknown	84				C	U				7.0.0
7670	*rho2*	transcription termination factor 2	458				C	C				2.2.11
7680	*hslV*	ATP-dependent protease HslV	189				C	C			3.4.25.-	2.1.4

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
7690	*hslU*	ATP-dependent hsl protease ATP-binding subunit	488				C	C				2.1.4
7700	*ubiE*	ubiquinone/menaquinone biosynthesis methyltransferase UbiE	230				C	U			2.1.1.-	3.2.08
7710	*metG*	methionyl-tRNA synthetase	506				C	C			6.1.1.10	2.2.01
7720		aspartate kinase*	415				C	U			2.7.2.4	3.1.13
7730	*coxB*	cytochrome c oxidase subunit 2*	258	3		2 th, signal	IM	IM			1.9.3.1	3.5.3
7740	*coxA*	cytochrome c oxidase subunit 1*	518	12			IM	IM			1.9.3.1	3.5.3
7750	*ctaB*	protoheme IX farnesyltransferase*	295	9		9 th	IM	IM			2.5.1.-	3.5.1
7760		exported lipoprotein*	250		signal	signal	C	U				4.1.0
7770	*purD*	phosphoribosylamine--glycine ligase	423				C	C			6.3.4.13	3.7.1
7780		preprotein translocase subunit YajC*	122	1		1 th ,signal	C	U				1.5.5
7790		unknown	234				OM	U				0.0.0
7800		outer membrane efflux protein*	415		signal	signal	OM	OM				1.5.5
7810	*rplM*	50S ribosomal protein L13	156				P	U				4.2.2
7820	*rpsI*	30S ribosomal protein S9	153				E	U				4.2.2
7830	*argC*	N-acetyl-gamma-glutamyl-phosphate reductase	347	1		1 th	C	U			1.2.1.38	3.1.02
7840	*ppa*	inorganic pyrophosphatase	173				C	C			3.6.1.1	3.3.13
7850		unknown	209				C	U	+			7.0.0
7860		response regulator component of a two-component regulatory system*	267				C	C				6.1.2
7870		exonuclease†	205				C	C				7.0.0
7880	*dnaN*	DNA polymerase III, beta subunit	375				C	C			2.7.7.7	2.2.03
7890		conserved hypothetical protein	349				C	U				0.0.2
7900	*prsA*	ribose-phosphate pyrophosphokinase	318				C	U			2.7.6.1	3.7.1
7910	*gatC*	glutamyl-tRNA(Gln) amidotransferase subunit C*	114				C	U			6.3.5.-	2.2.01
7920	*acnA*	aconitate hydratase	875				OM	C			4.2.1.3	3.5.8
7930		conserved hypothetical protein	134				C	U				0.0.2
7940	*purK*	phosphoribosylaminoimidazole carboxylase ATPase subunit	359				C	U			4.1.1.21	3.7.1
7950		ATP/GTP-binding membrane protein*	735	1		1 th	OM	U				4.1.1
7960		unknown	1304			signal	OM	OM		15 bp x 2.7		7.0.0
7970		exported protein*	1710		signal	signal	OM	OM				4.1.0
7980		type IV secretion system protein†	790				C	C				1.5.6
7990		integral membrane protein*	125	3		signal, 2 th	IM	IM				4.1.1
8000		integral membrane protein*	112	3		signal, 2 th	IM	IM				4.1.1
8010		integral membrane protein*	118	3		signal, 2 th	IM	IM				4.1.1
8020		integral membrane protein*	124	3		signal, 2 th	IM	IM				4.1.1
8030	*hflK*	HflK protein†	356			1 th	C	U				7.0.0
8040	*hflC*	HflC membrane protein†	290	1‡		1 th	C	U			3.4.-.-	4.1.1
8050		exported serine protease*	476		signal	signal	OM	P			3.4.21.-	2.1.4
8060		exported protein*	204		signal	signal	OM	U				4.1.0

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
8070	*rnc*	ribonuclease III	227				C	C			3.1.26.3	2.1.2
8080	*ctaG*	cytochrome c oxidase assembly protein	177		signal	signal	P	U				2.2.13
8090		exported peptidase*	438		signal	signal	C	P			3.4.24.-	2.1.4
8100		exported M16 family peptidase*	455		signal	signal	E	U			3.4.24.-	2.1.4
8110		integral membrane protein*	224	4		4 th	IM	IM				4.1.1
8120	*lspA*	lipoprotein signal peptidase	149	3		4 th	IM	IM			3.4.23.36	2.2.06
8130	*ribF*	riboflavin kinase/FAD synthetase	306				C	C			2.7.1.26, 2.7.7.2	3.2.13
8140	*grxC*	glutaredoxin 3*	95				C	U				3.2.15
8150		methyltransferase*	280				C	U			2.1.1.-	7.0.0
8160	*map*	methionine aminopeptidase	266				C	C			3.4.11.18	2.2.10
8170		unknown	372				C	IM				0.0.0
8180		unknown	150				OM	U				0.0.0
8190		unknown	563				OM	C				0.0.0
8200	*sucB*	dihydrolipoamide succinyltransferase, E2 component of 2-oxoglutarate dehydrogenase complex	402				C	C			2.3.1.61	3.5.8
8210		transferase†	172				C	U				7.0.0
8220		exported D-alanyl-D-alanine carboxypeptidase*	290		signal	signal	IM	U			3.4.16.4	2.1.4
8230		integral membrane protein*	402	8		8 th	IM	IM				4.1.1
8240		conserved hypothetical protein	93				C	C				0.0.1
8250		membrane-associated zinc metalloprotease*	379	4		4 th	IM	IM			3.4.24.-	2.1.4
8260		outer membrane protein*	771		signal	signal	OM	OM				4.1.4
8270		outer membrane protein*	182		signal	signal	C	U				4.1.3
8280	*fabZ*	(3R)-hydroxymyristoyl-[acyl carrier protein] dehydratase	145				C	U			4.2.1.-	3.6.1
8290	*purH*	bifunctional purine biosynthesis protein PurH	504				C	U			3.5.4.10, 2.1.2.3	3.7.1
8300	*pgsA*	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase*	185	5		6 th	IM	IM			2.7.8.5	2.2.07
8310		integral membrane protein*	433	12		12 th	IM	IM				4.1.1
8320		Surf1-like protein†	213	2		2 th	IM	IM				4.1.1
8330		conserved hypothetical protein	609		signal	signal	E	OM				0.0.2
8340		unknown	622				OM	U				0.0.0
8350		unknown	419			2 th	C	C				7.0.0
8360	*atpB*	ATP synthase A subunit	243	7		7 th	IM	IM			3.6.3.14	3.5.9
8370	*atpE*	ATP synthase C subunit	73	2		2 th	IM	U			3.6.3.14	3.5.9
8380	*atpF*	ATP synthase B subunit*	167	1		1 th	C	U			3.6.3.14	3.5.9
8390		membrane protein*	163	1		1 th	C	C				4.1.1
8400	*ftsA*	cell division protein FtsA	419				C	U				1.7.1
8410	*trkH*	Trk system potassium uptake protein	483	12		12 th	IM	IM				1.5.2
8420		conserved hypothetical protein	442				C	C				0.0.2

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
8430	*ftsH*	cell division protein FtsH	611	2‡		1 th, signal	IM	IM			3.4.24.-	2.1.4
8440	*lgt*	prolipoprotein diacylglyceryl transferase	259	7		7 th	IM	IM			2.4.99.-	2.2.07
8450		integral membrane protein*	560	5		5 th	OM	IM		243 bp x 5.2 (C-terminus)		4.1.1
8460		unknown	56				C	C				0.0.0
8470	*secD*	protein-export membrane protein SecD*	505	5		5 th	IM	IM				1.5.5
8480	*thiF*	adenylyltransferase ThiF*	260	1‡			OM	IM			2.7.7.-	3.2.14
8490	*pyrE*	orotate phosphoribosyltransferase*	199				C	U			2.4.2.10	3.7.2
8500	*recA*	RecA protein (Recombinase A)	357				C	U				2.2.03
8510		membrane protein*	223	1			C	U			6.3.3.3	4.1.1
8520	*ftsY*	cell division protein FtsY*	309				C	C				1.7.1
8530	*icd*	isocitrate dehydrogenase [NADP]	483				C	C			1.1.1.42	3.5.8
8550	*recJ*	single-stranded-DNA-specific exonuclease RecJ*	585			2 th	C	C			3.1.-.-	2.1.1
8560		nucleic acid independent RNA polymerase*	397				C	U			2.7.7.-	2.2.11
8570	*ndk*	nucleoside diphosphate kinase	143				C	C			2.7.4.6	3.3.11
8580		transcriptional regulator†	201				C	U	+			6.1.2
8590	*map1-14*	outer membrane protein MAP1-14*	309		signal	signal	P	IM		6 bp x 3.7		4.1.3
8600	*map1-13*	outer membrane protein MAP1-13*	294		signal	signal	OM	IM				4.1.3
8610		exported protein*	236		signal	signal	OM	U				4.1.0
8620	*map1-12*	outer membrane protein MAP1-12*	275		signal	signal	OM	IM				4.1.3
8630	*map1-11*	outer membrane protein MAP1-11*	293		signal	signal	OM	U				4.1.3
8640	*map1-10*	outer membrane protein MAP1-10*	257		signal	signal	OM	U				4.1.3
8650	*map1-9*	outer membrane protein MAP1-9*	289		signal	signal	OM	IM				4.1.3
8660	*map1-8*	outer membrane protein MAP1-8*	282		signal	signal	OM	U				4.1.3
8670	*map1-7*	outer membrane protein MAP1-7*	283		signal	signal	OM	U				4.1.3
8680	*map1-6*	outer membrane protein MAP1-6*	295		signal	signal	OM	OM				4.1.3
8690	*map1-5*	truncated outer membrane protein MAP1-5†	205				OM	U				4.1.3
8700	*map1-4*	outer membrane protein MAP1-4*	297		signal	signal	OM	IM				4.1.3
8710	*map1-3*	outer membrane protein MAP1-3*	315		signal	signal	E	U				4.1.3
8720	*map1-2*	outer membrane protein MAP1-2*	306		signal	signal	OM	U				4.1.3
8730	*map1-1*	outer membrane protein MAP1-1*	282		signal	signal	OM	U				4.1.3
8740	*map1*	major antigenic protein MAP1	290		signal	signal	E	U				4.1.4
8750	*map1+1*	outer membrane protein MAP1+1*	285			signal	OM	U				4.1.3
8760		unknown	111				C	U				0.0.0
8770		unknown	177				C	C		24 bp x 6.3		0.0.0
8780	*secA*	preprotein translocase SecA subunit	870				C	C				1.5.5
8790		unknown	143				C	U				0.0.0
8800	*ftsZ*	cell division protein FtsZ	422				C	U				1.7.1
8810		conserved hypothetical protein	135				C	U				0.0.2

Erum ID	gene name	product	length (aa)	TMHMM	SignalP	Phobius	CELLO	pSORTb	HTH	tandem repeats	EC number	class
8820		conserved hypothetical protein	157				C	U				0.0.2
8830	*parA*	chromosome partitioning protein ParA	255				OM	U				1.2.1
8840	*parB*	chromosome partitioning protein ParB	289				C	U	+			1.2.1
8850	*rimM*	16S rRNA processing protein*	172				C	U				2.2.11
8860	*trmD*	tRNA (Guanine-N(1)-)-methyltransferase	235				OM	U			2.1.1.31	2.2.11
8870	*rplS*	50S ribosomal protein L19	125				C	C				4.2.2
8880		unknown	166				C	U				7.0.0
8890	*thrS*	threonyl-tRNA synthetase	633				C	C			6.1.1.3	2.2.01
8900	*infC*	translation initiation factor IF-3	173				C	C				2.2.10
8910		conserved hypothetical protein	150				C	U				0.0.2
8920		integral membrane protein*	234	6		7 th	IM	IM				4.1.1
8930		integral membrane protein*	409	6		6 th	IM	IM				4.1.1

*probable; †possible; ‡The initial transmembrane helix could represent a possible N-terminal signal sequence.

Appendix F: Web based tools

Web based tools used in this study

Annotation

BioCyc	http://biocyc.org/ecocyc/index.shtml
KEGG	http://www.genome.jp/kegg/pathway.html
Pfam	http://pfam.sanger.ac.uk/
PROSITE	http://www.expasy.ch/prosite/
Tandem Repeats Finder	http://tandem.bu.edu/trf/trf.html

Subcellular localisation

CELLO	http://cello.life.nctu.edu.tw/
Phobius	http://phobius.cgb.ki.se/
PSORTb v.2.0	http://www.psort.org/psortb/
SignalP	http://www.cbs.dtu.dk/services/SignalP/
TMHMM2.0	http://www.cbs.dtu.dk/services/TMHMM-2.0/

Recombinant protein analysis

Protein Molecular Weight	http://www.bioinformatics.org/sms/prot_mw.html
Recombinant Protein Solubility Prediction	http://biotech.ou.edu/

Appendix G: Publications and ethics

G1: Publications

The research conducted in this study has been published in the following articles:

COLLINS, N.E., LIEBENBERG, J., DE VILLIERS, E.P., BRAYTON, K.A., LOUW, E., PRETORIUS, A., FABER, F.E., VAN HEERDEN, H., JOSEMANS, A., VAN KLEEF, M., *et al.* 2005. The genome of the heartwater agent *Ehrlichia ruminantium* contains multiple tandem repeats of actively variable copy number. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 838-843.

PRETORIUS, A., LIEBENBERG, J., LOUW, E., COLLINS, N.E. & ALLSOPP B.A. 2010. Studies of a polymorphic *E. ruminantium* gene for use as a component of a recombinant vaccine against heartwater. *Vaccine* **28**: 3531-3539.

SEBATJANE, S.I., PRETORIUS, A., LIEBENBERG, J. STEYN, H.C., & VAN KLEEF, M. 2010. *In vitro* and *in vivo* evaluation of five low molecular weight proteins of *Ehrlichia ruminantium* as potential vaccine components. *Veterinary Immunology and Immunopathology* **137**: 217-225.

Article in preparation for publication (Chapter 5):

LIEBENBERG, J., PRETORIUS, A., FABER, F.E., J. HEATH, J., COLLINS, N.E., VAN KLEEF, M. & ALLSOPP, B.A. Identification of novel potential vaccine candidates against *Ehrlichia ruminantium*. To be submitted to *Veterinary microbiology.*

G2: Ethics

The research presented in this thesis was approved by the Animal Ethics committee of the ARC-Onderstepoort Veterinary Institute and the Animal Use and Care committee of the University of Pretoria (protocol V036/06).