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°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.

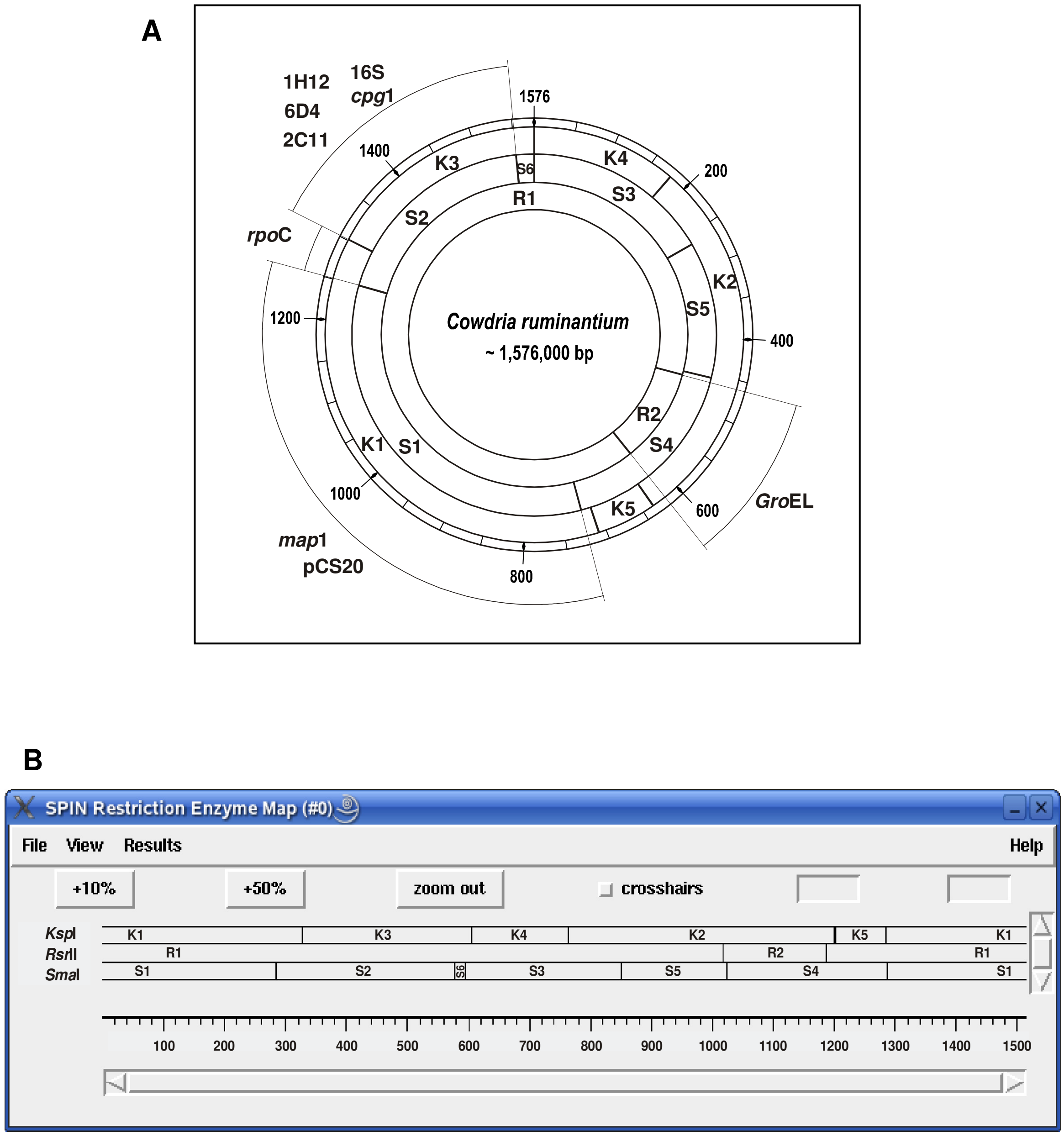


Figure 2.1. **A.** The physical map of De Villiers *et al*. (2000). From inside to outside, the circles represent the *Rsr*II, *Sma*I and *Ksp*I restriction sites respectively. The outer circle illustrates the scale in kilobases. Mapped genes and clones are also indicated. **B.** A computer-generated 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*).

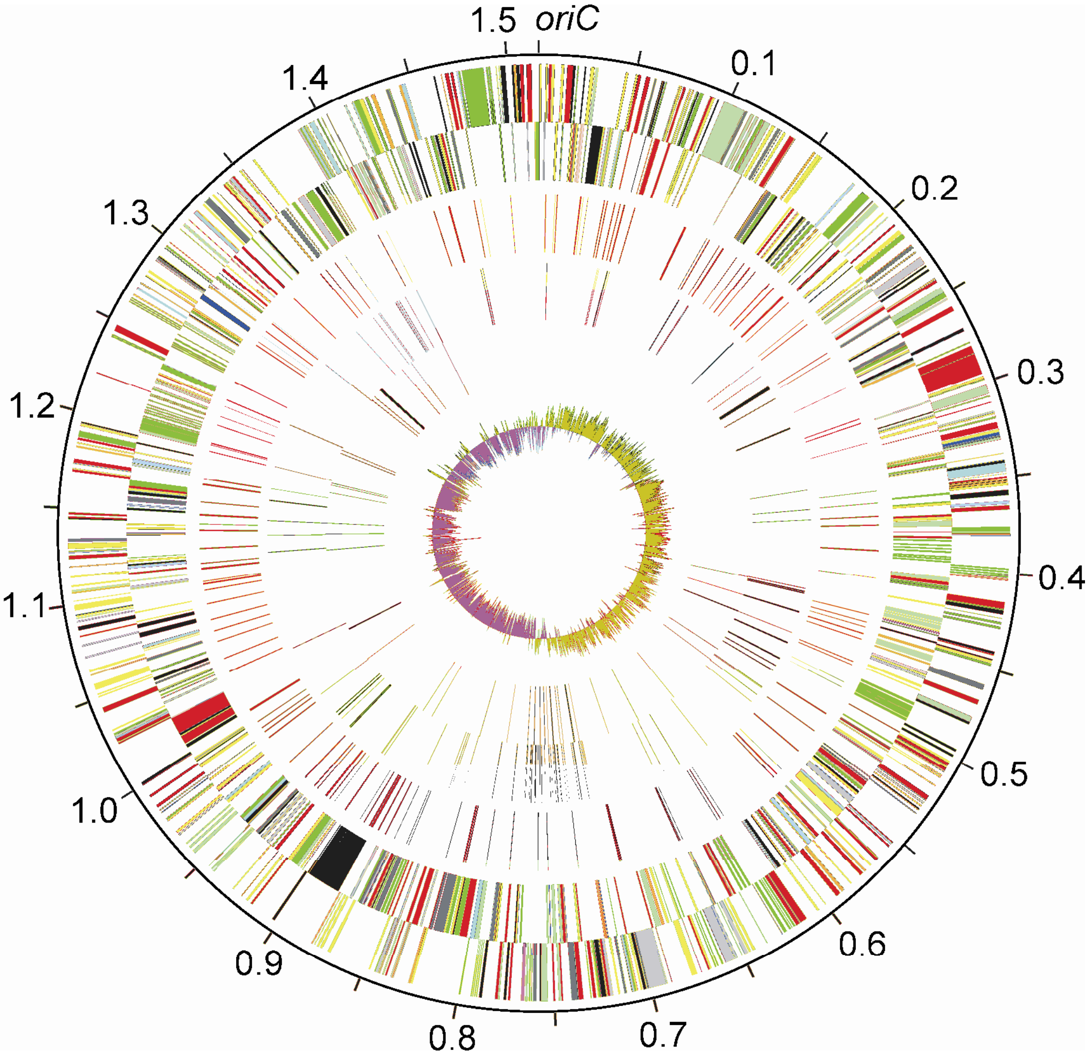
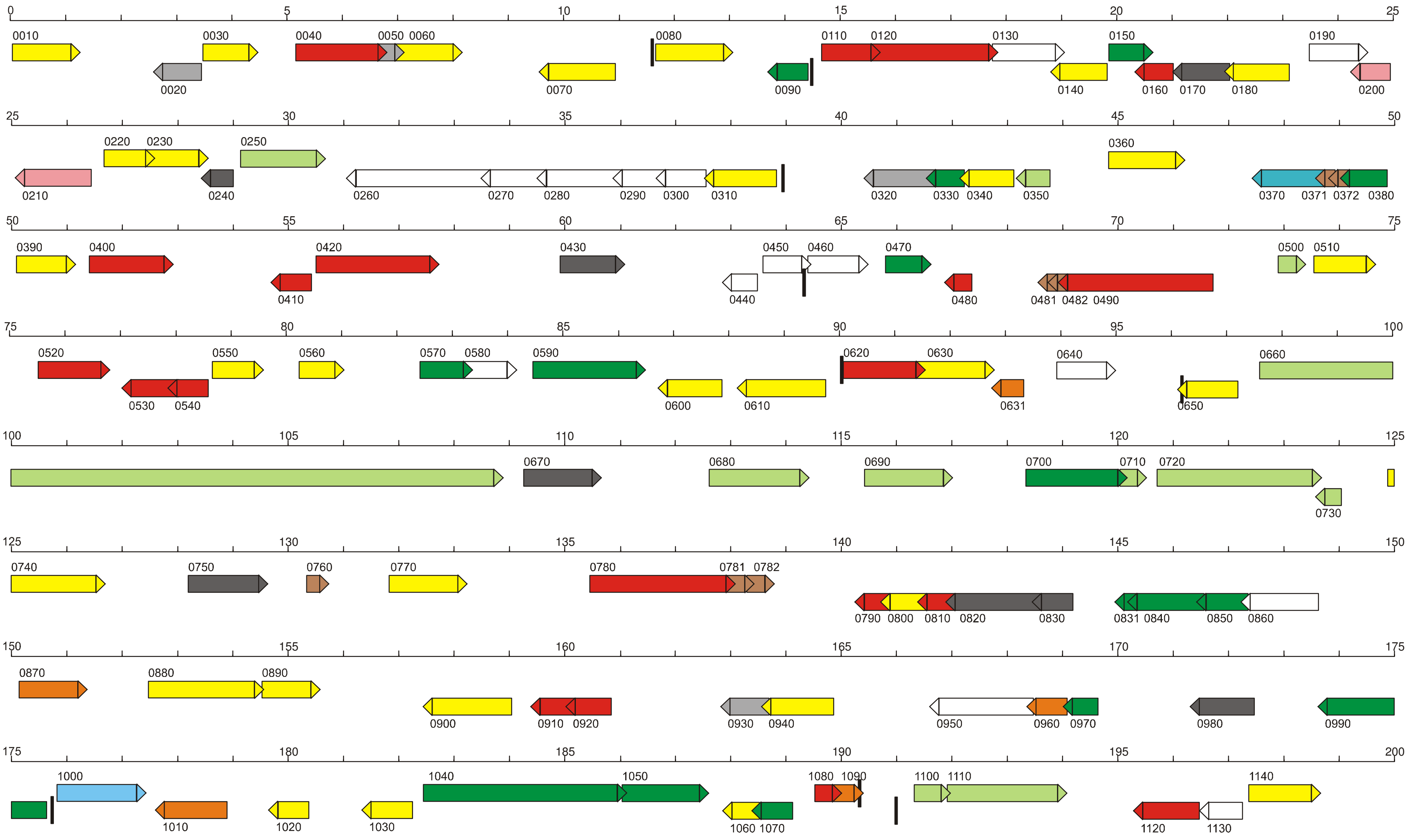
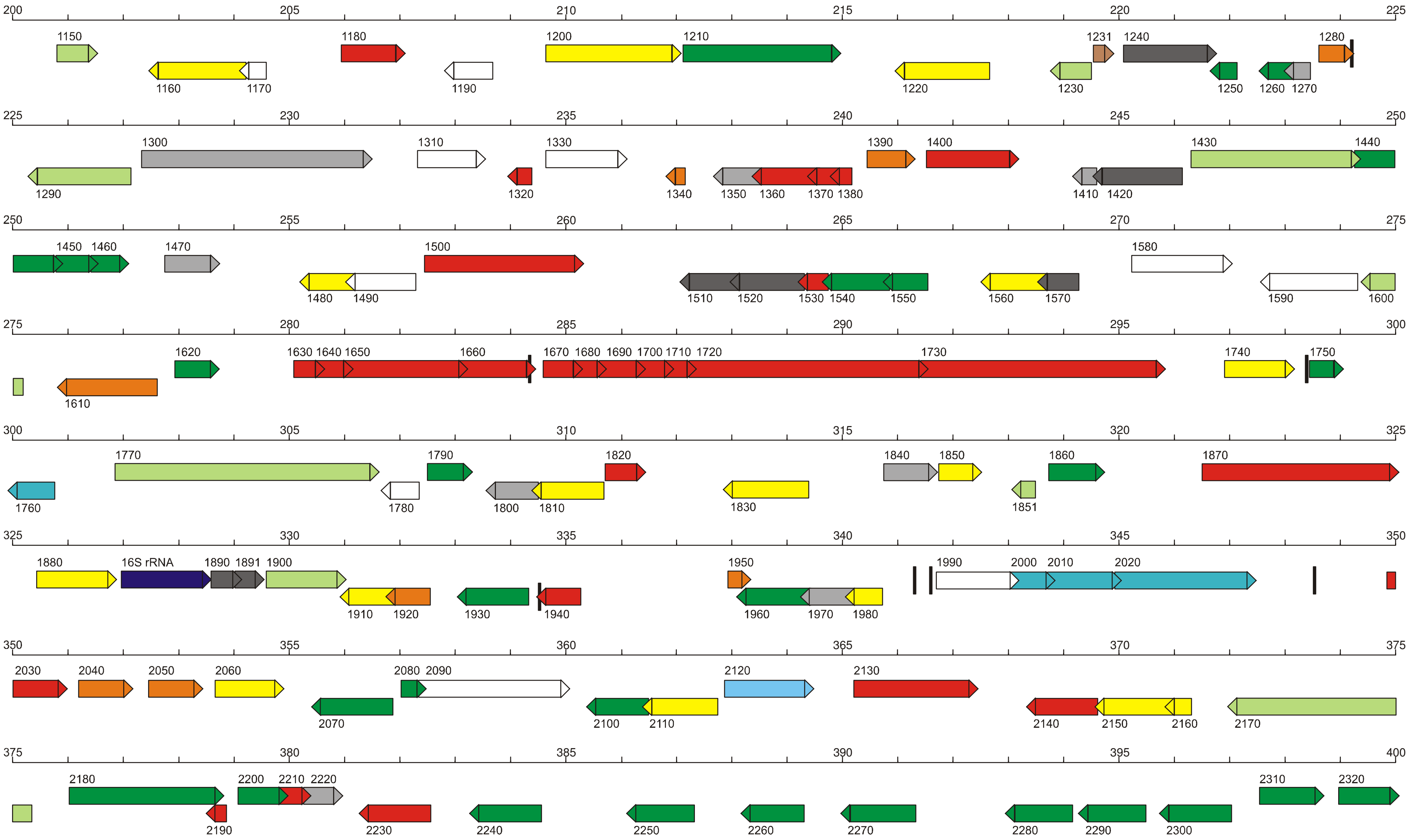


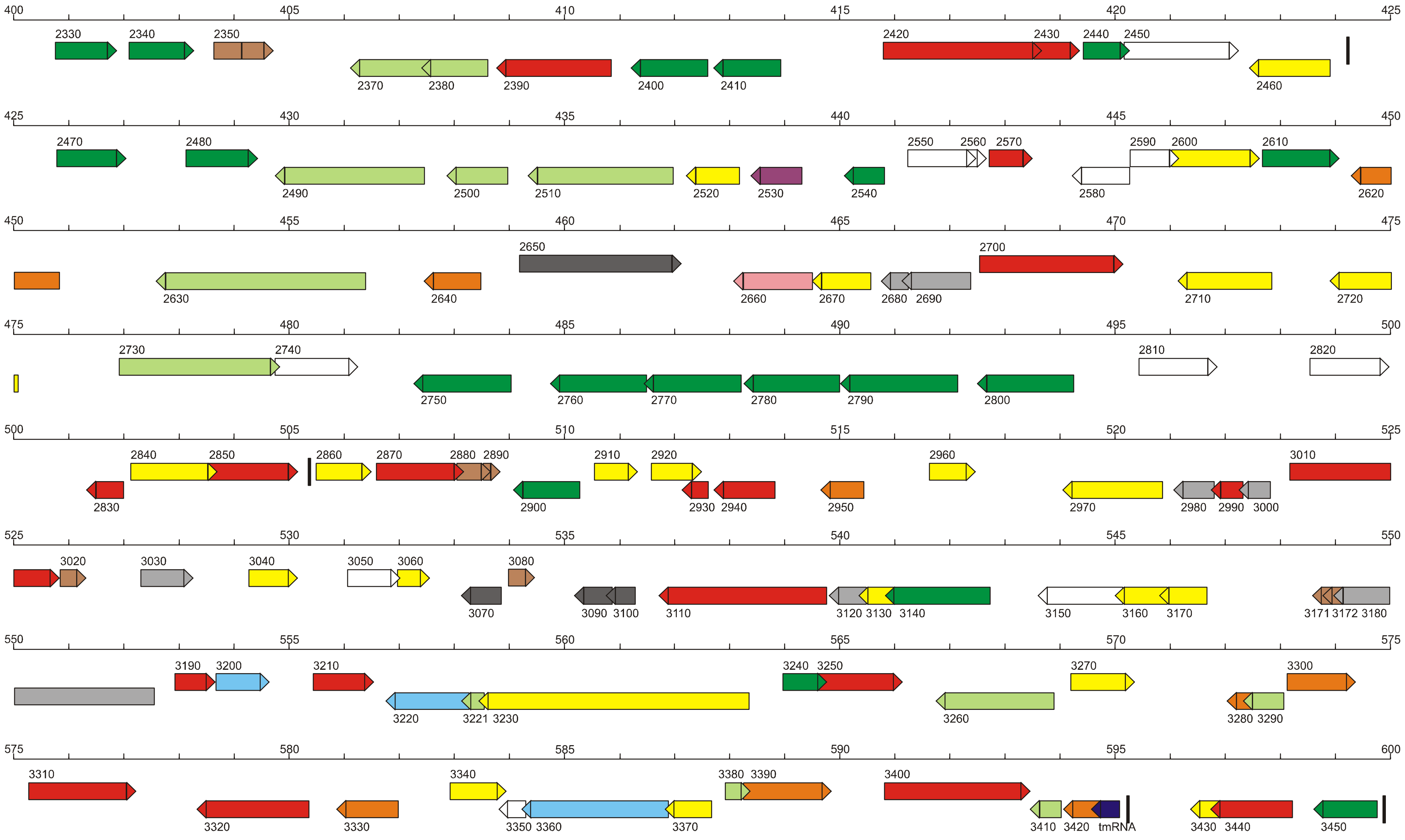
Figure 2.2. Circular representation of the genome of *E. ruminantium* (Welgevonden isolate). The outer circle indicates the scale in megabases. The remaining concentric circles are described from outside to inside. First and second circles, predicted coding sequences on the plus and minus strands respectively, 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.)

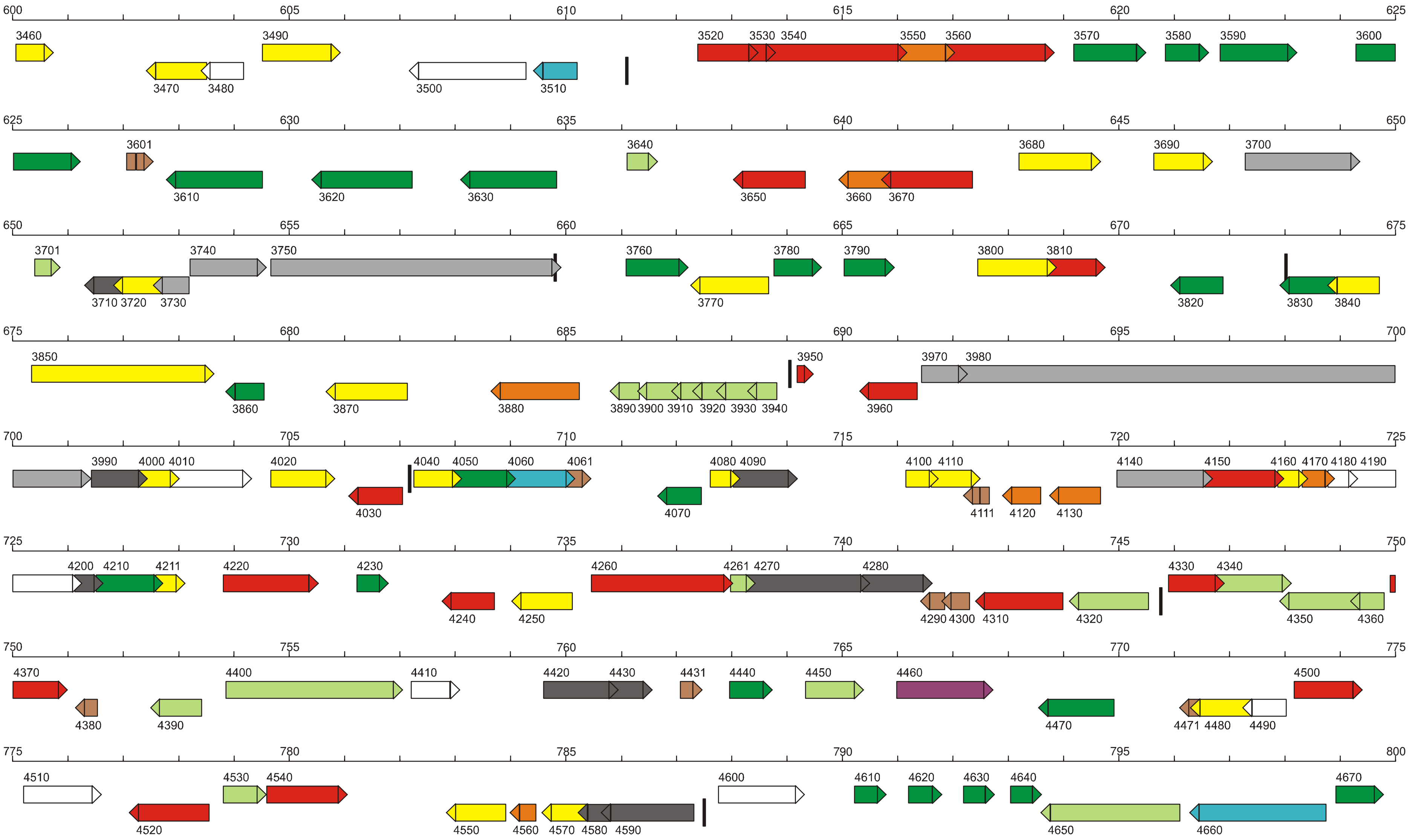
The origin of replication (*oriC*) has not been experimentally determined in *E. ruminantium*. In many other organisms there is a conserved arrangement of genes around *oriC* (Ogasawara & Yoshikawa, 1992), which is often located close to the *dnaA* gene, and a transition in GC-skew values is frequently 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.

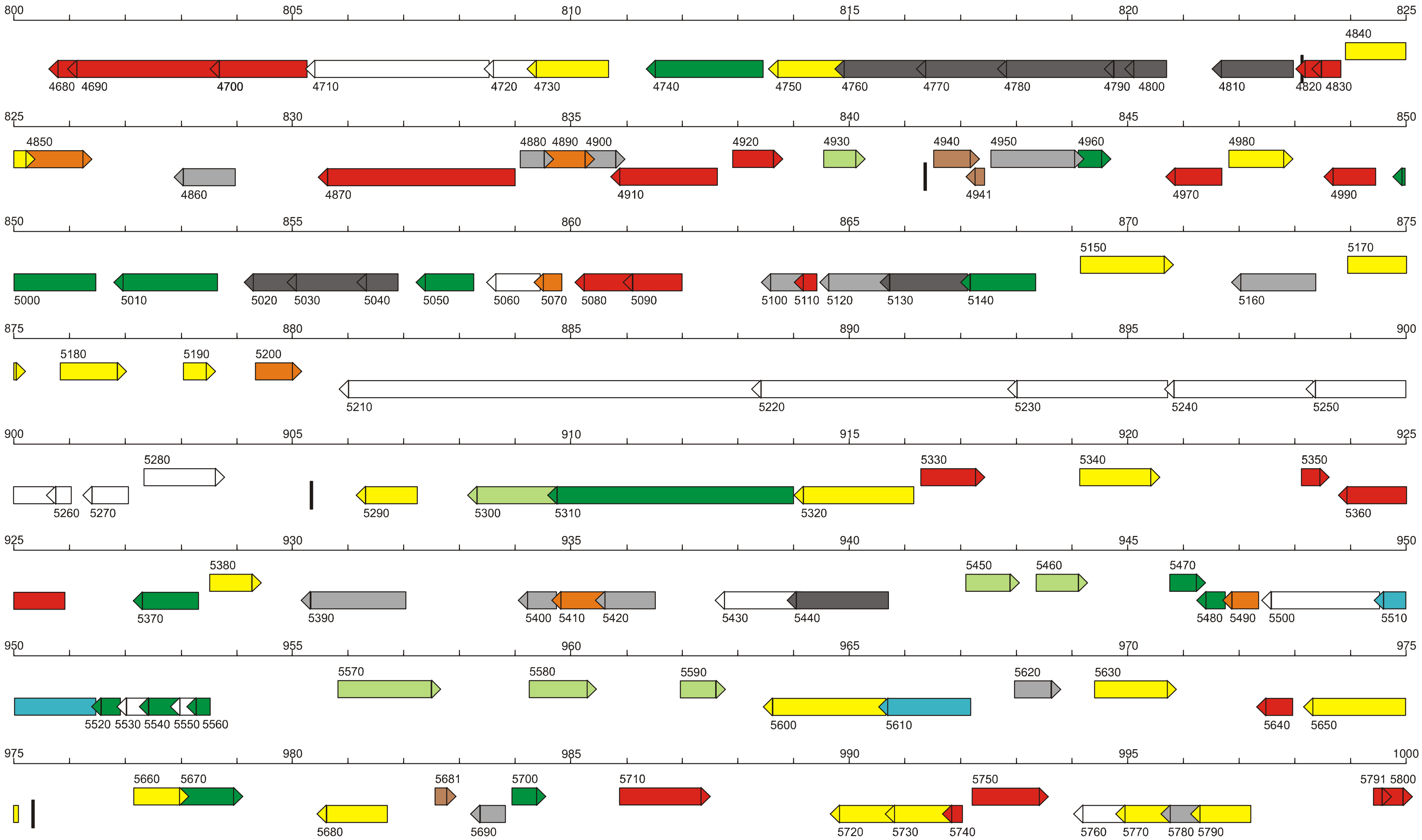
The unusual dispersion in *E. ruminantium* of genes normally found to be associated with *oriC* was also observed with other genes that normally occur in operons in other bacteria. One such example is the disruption of the ribosomal RNA (rRNA) operon: the 16S rRNA gene is located at 326,964 bp while 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.

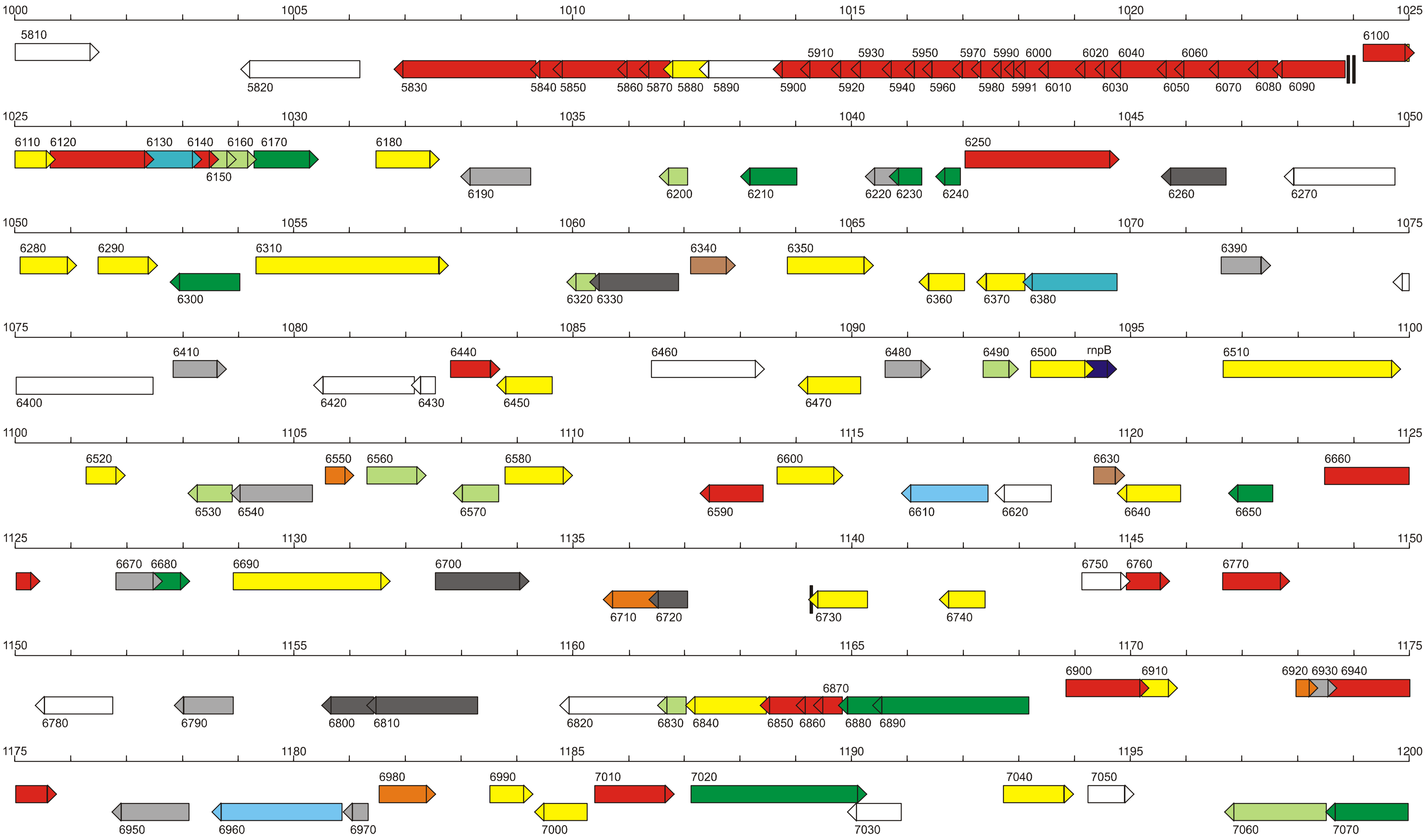

0
5
10
15
20
25
0010
0030
0040
0050
0060
0080
0110
0120
0130
0150
0190
0020
0070
0090
0140
0160
0170
0180
0200
25
30
35
40
45
50
0220
0230
0250
0360
0210
0240
0260
0270
0280
0290
0300
0310
0320
0330
0340
0350
0370
0371
0372
0380
50
55
60
65
70
75
0390
0400
0420
0430
0450
0460
0470
0500
0510
0410
0440
0480
0481
0482
0490
75
80
85
90
95
100
0520
0550
0560
0570
0580
0590
0620
0630
0640
0660
0530
0540
0600
0610
0631
0650
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
1040
1050
1080
1090
1100
1110
1140
1010
1020
1030
1060
1070
1120
1130

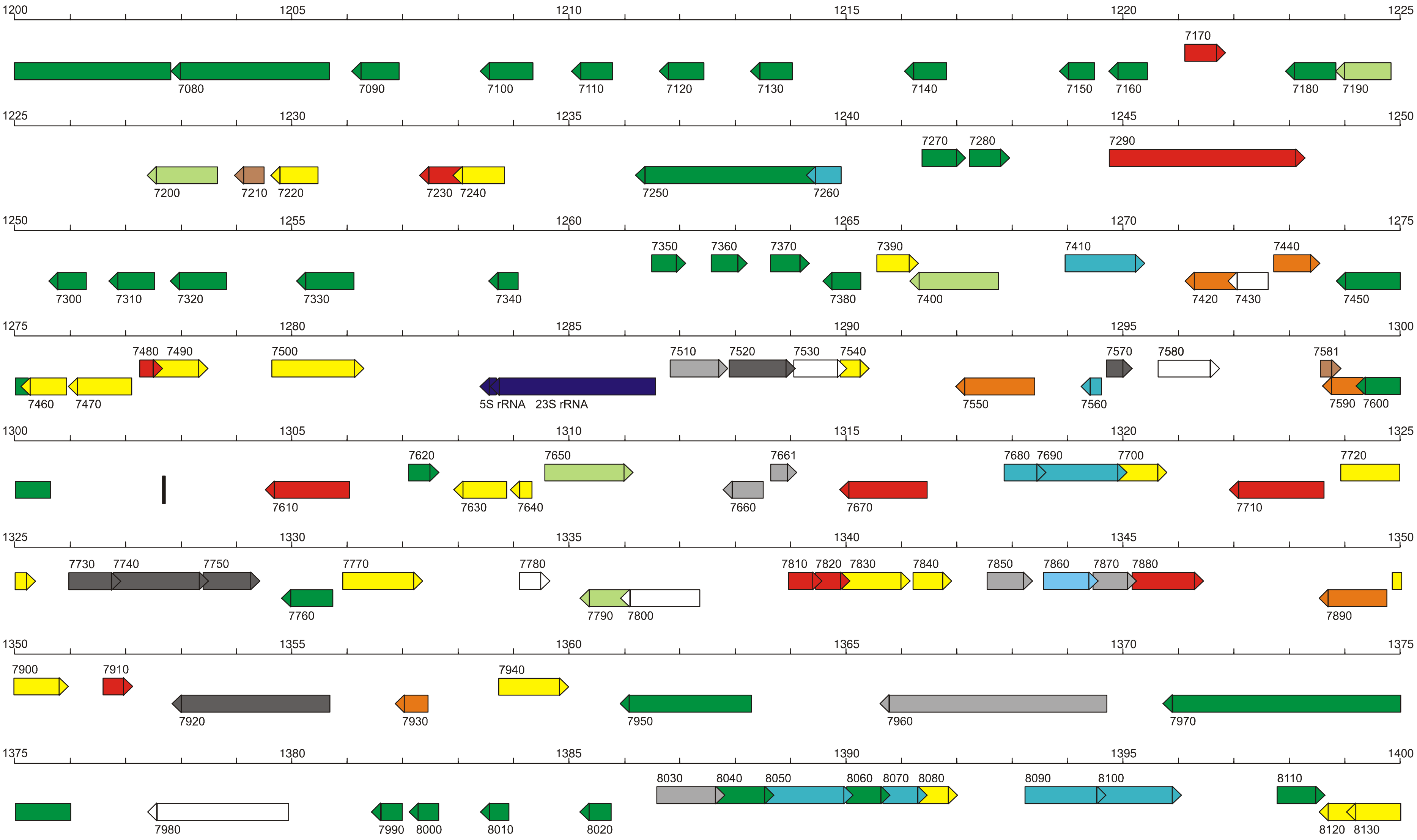

200
205
210
215
220
225
1150
1180
1200
1210
1231
1240
1280
1160
1170
1190
1220
1230
1250
1260
1270
225
230
235
240
245
250
1300
1310
1330
1390
1400
1430
1440
1290
1320
1340
1350
1360
1370
1380
1410
1420
250
255
260
265
270
275
1450
1460
1470
1500
1580
1480
1490
1510
1520
1530
1540
1550
1560
1570
1590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


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


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


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


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

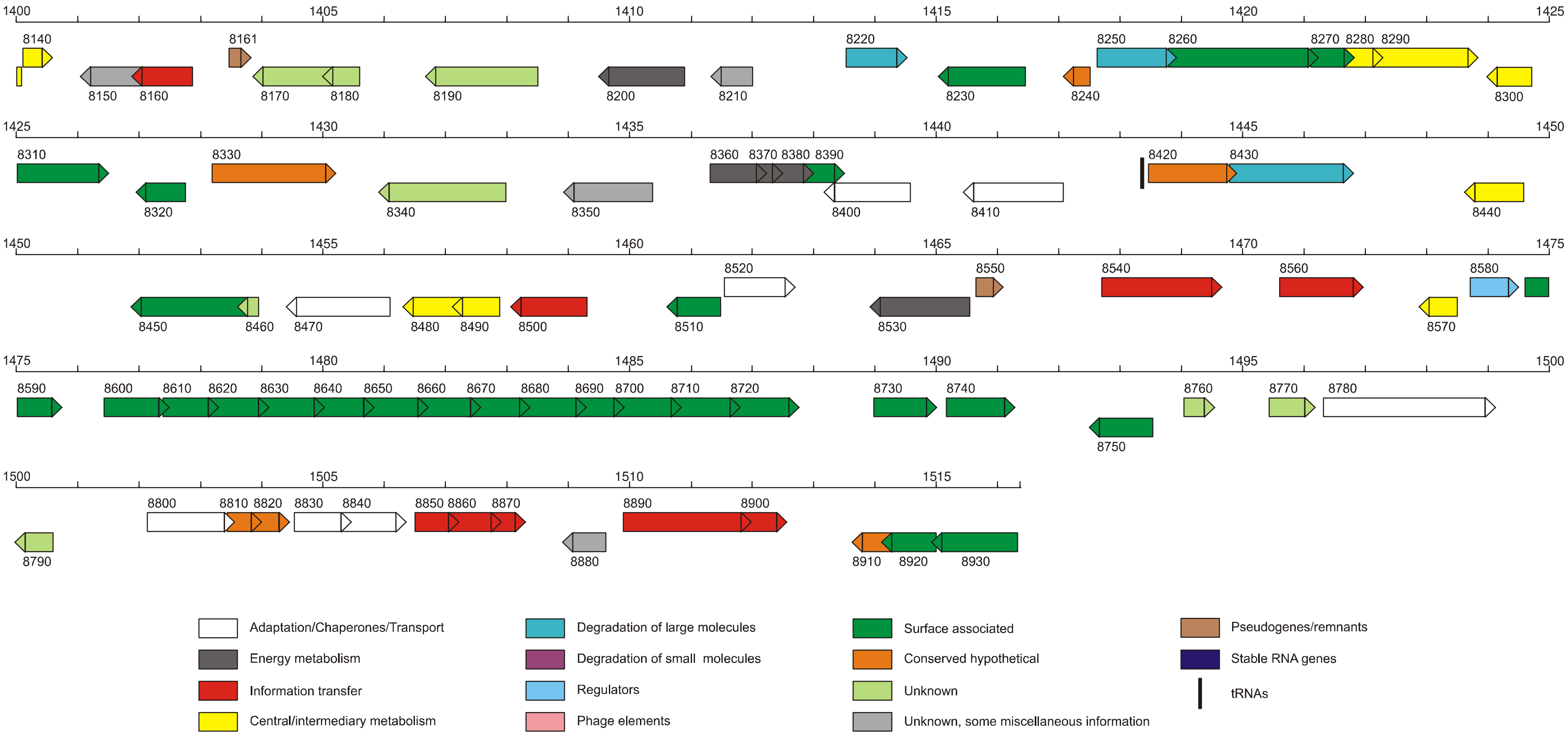


Figure 2.3. (Above and previous seven pages.) Linear representation of the *E. ruminantium* (Welgevonden isolate) genome. The scale of the map is shown in 1-kb increments. The potential protein coding regions (colour-coded by biological role) are depicted as boxes with arrowheads indicating the direction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

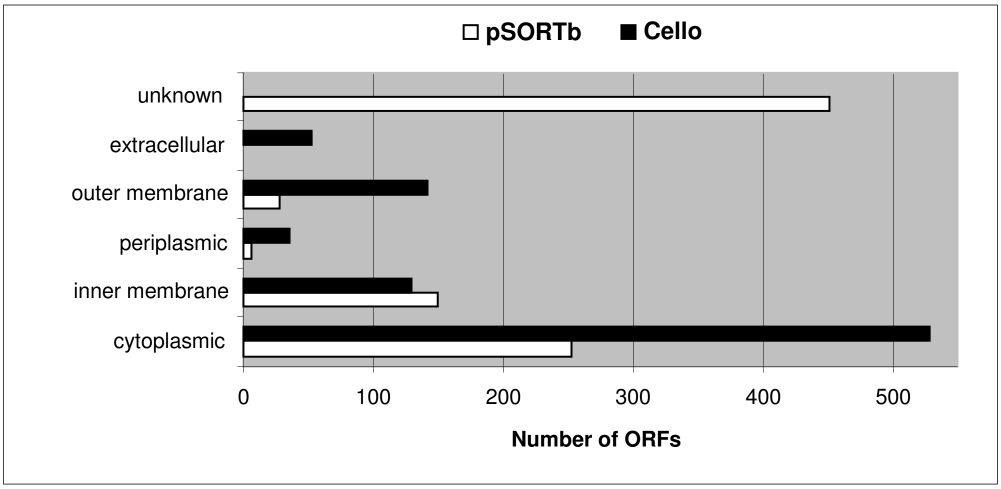


Figure 2.4. Predicted compartmentalisation of putative proteins by pSORTb and CELLO.

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

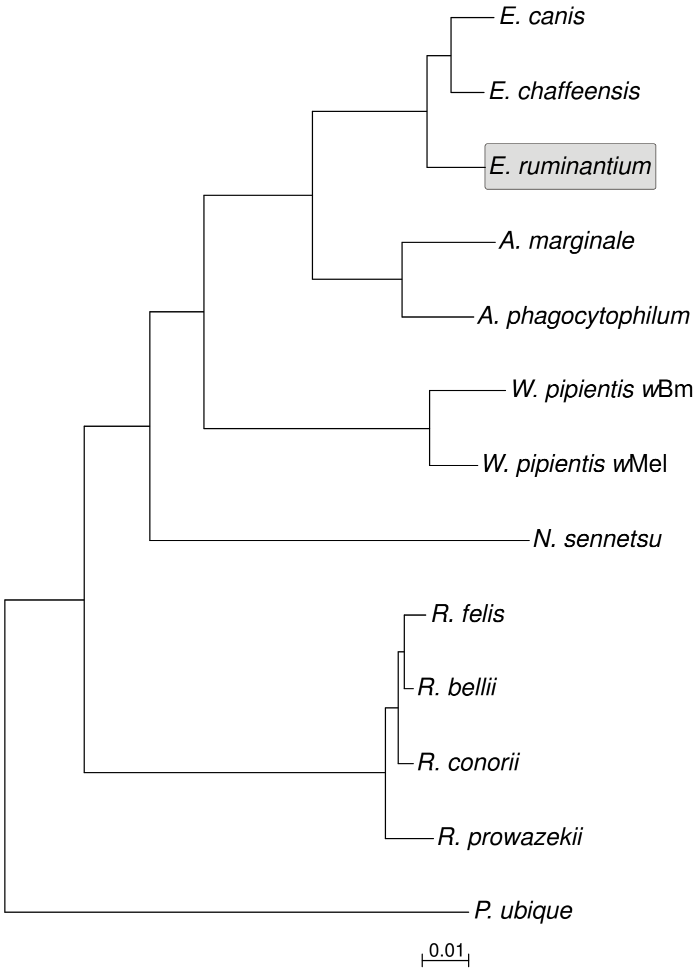


Figure 3.1. Neighbour-joining tree based on 16S rRNA sequences showing the phylogenetic relationships of *E. ruminantium* with other Rickettsiales for which complete genome sequences had been published at the time of this study. The sequences were aligned using ClustalX (Thompson *et al*., 2002) and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

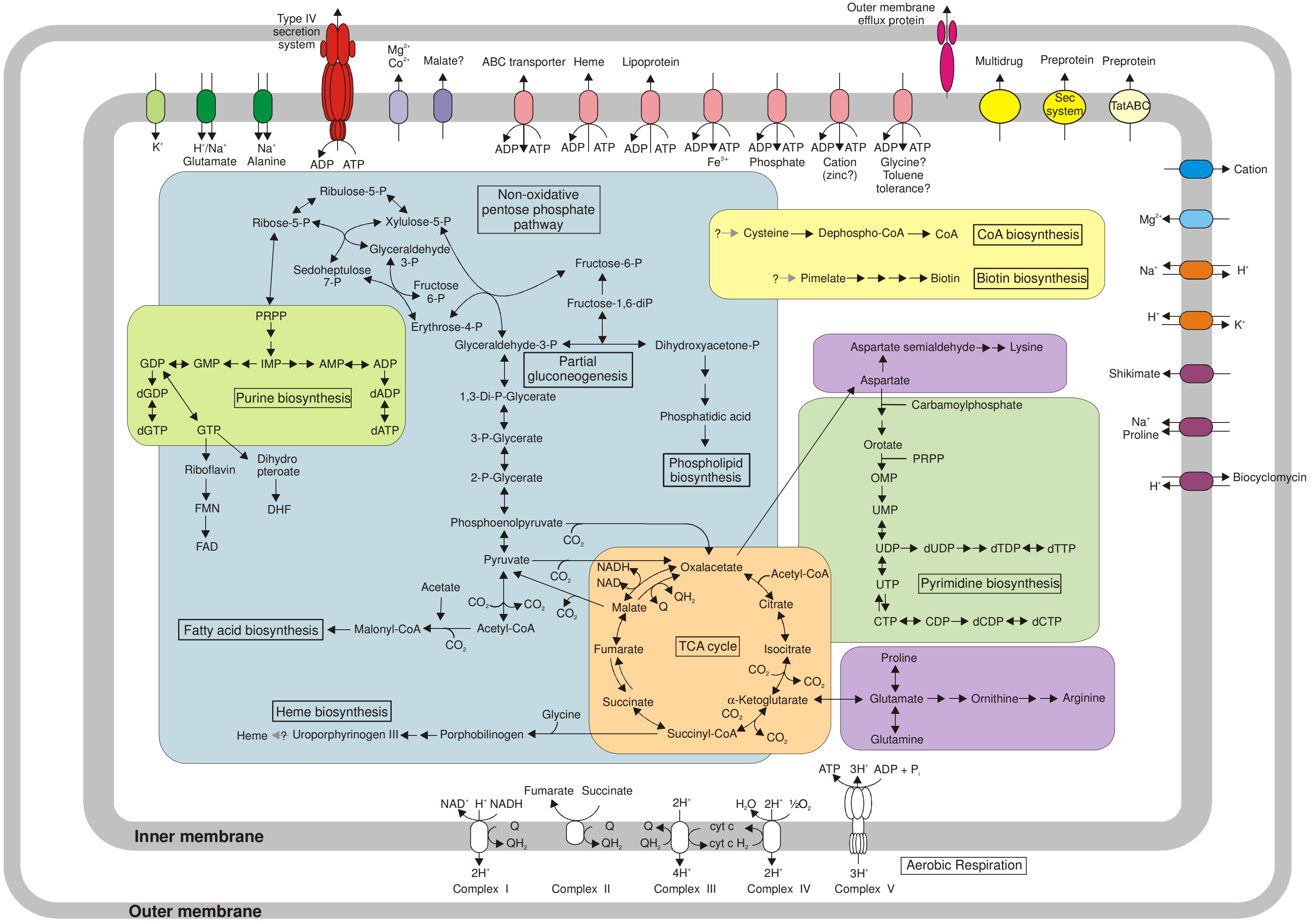


Figure 3.2. Schematic overview of metabolic pathways and substrate transport in *E. ruminantium*. Uncertainties are denoted by question marks. (Adapted from Collins *et al.*, 2005. [Supplementary information]).

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).

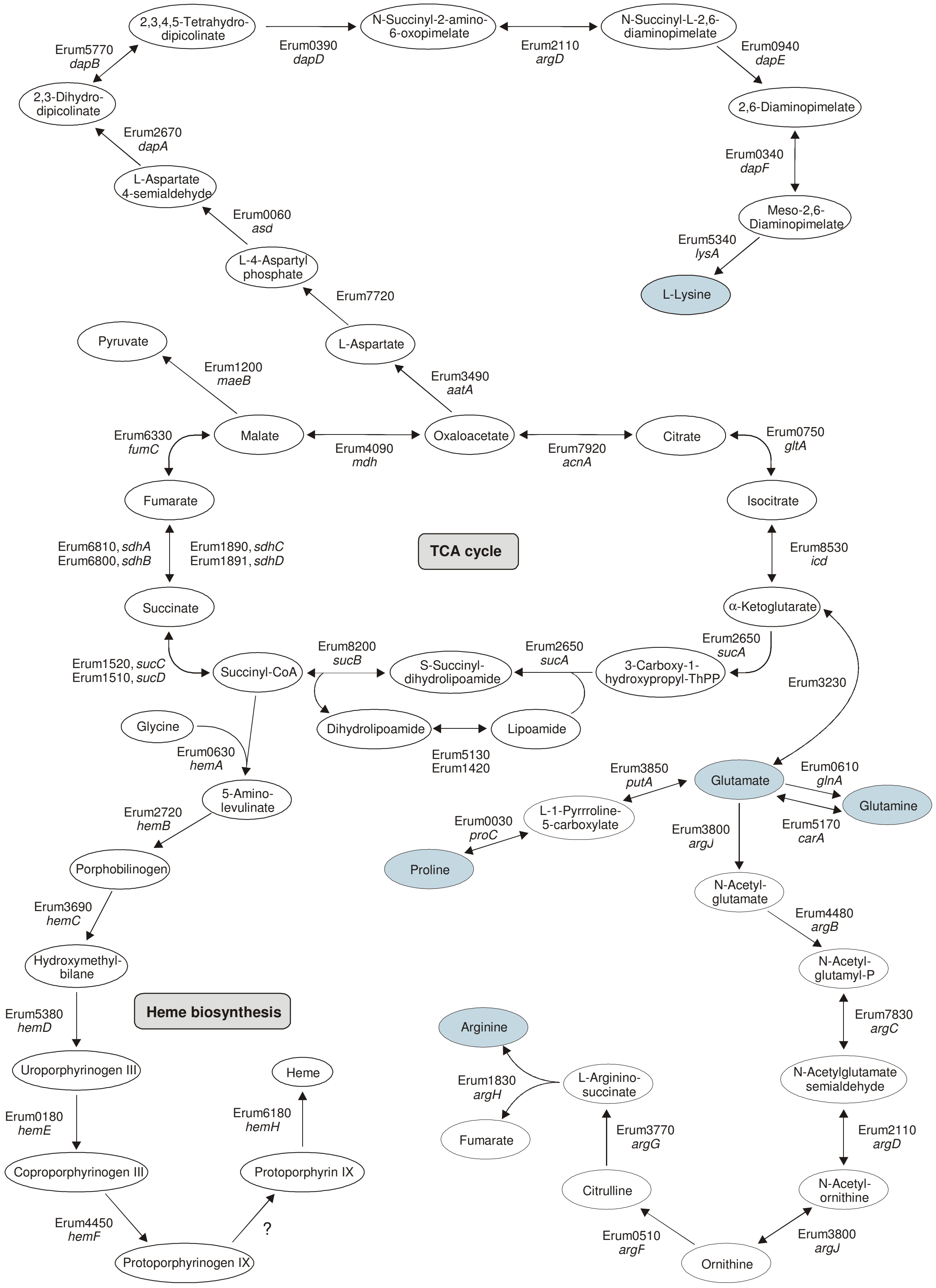


Figure 3.3. *E. ruminantium* genes coding for the enzymes involved in the TCA cycle, heme biosynthesis and amino acid biosynthesis. Amino acids for which pathways were identified are indicated in grey ovals.

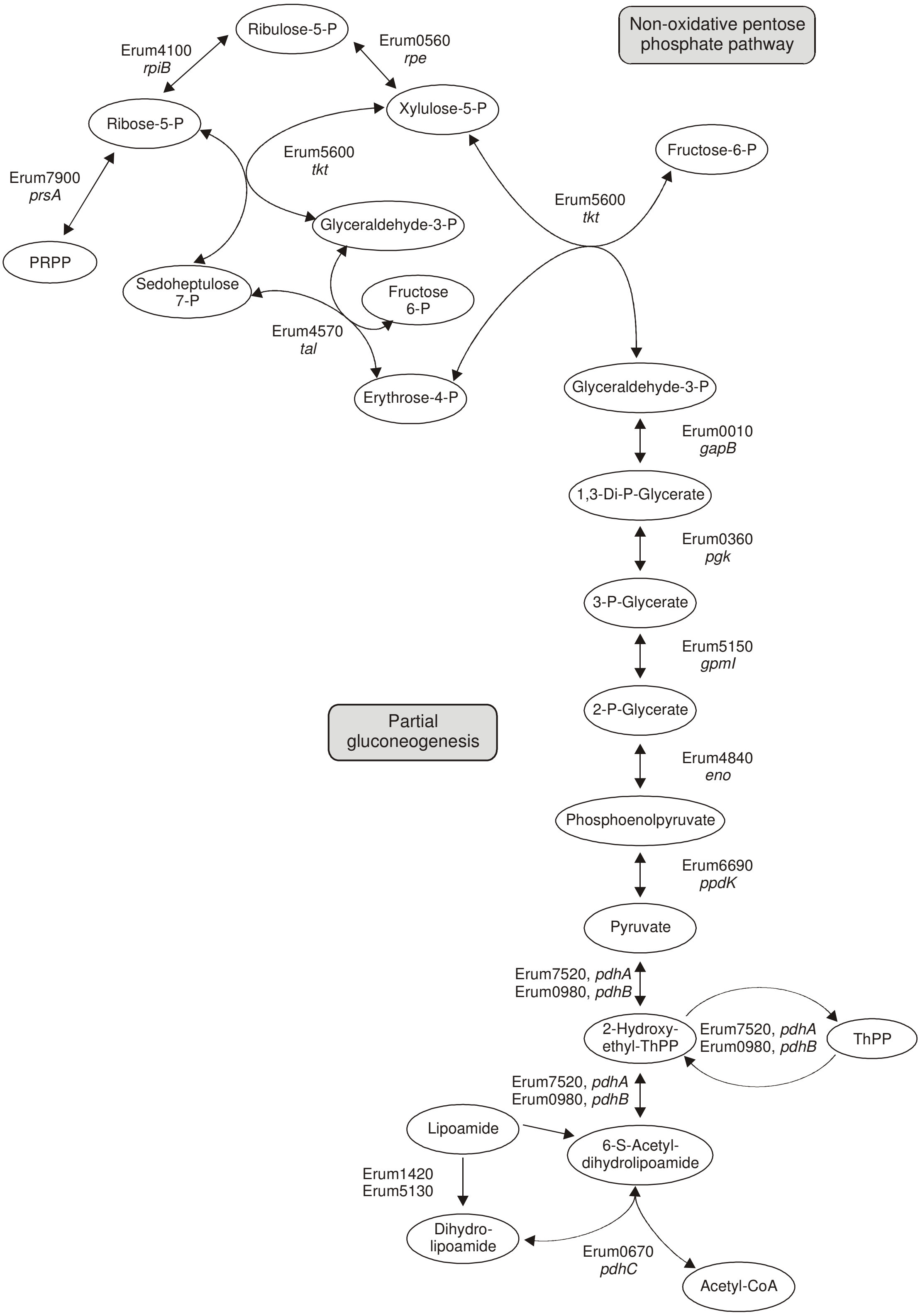


Figure 3.4. *E. ruminantium* genes involved in the pentose phosphate and gluconeogenesis pathways.

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).

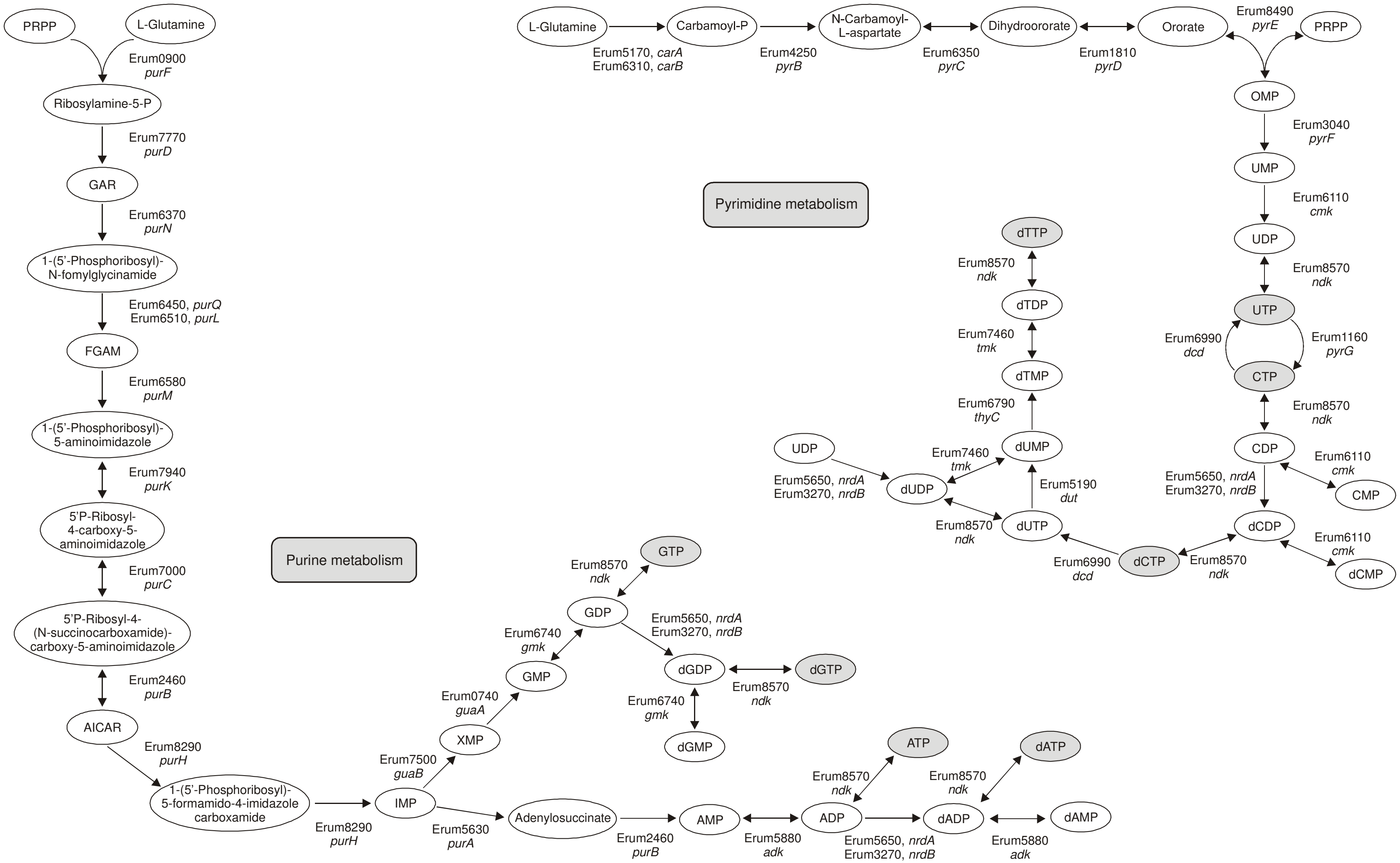


Figure 3.5. *E. ruminantium* genes involved in nucleotide metabolism.

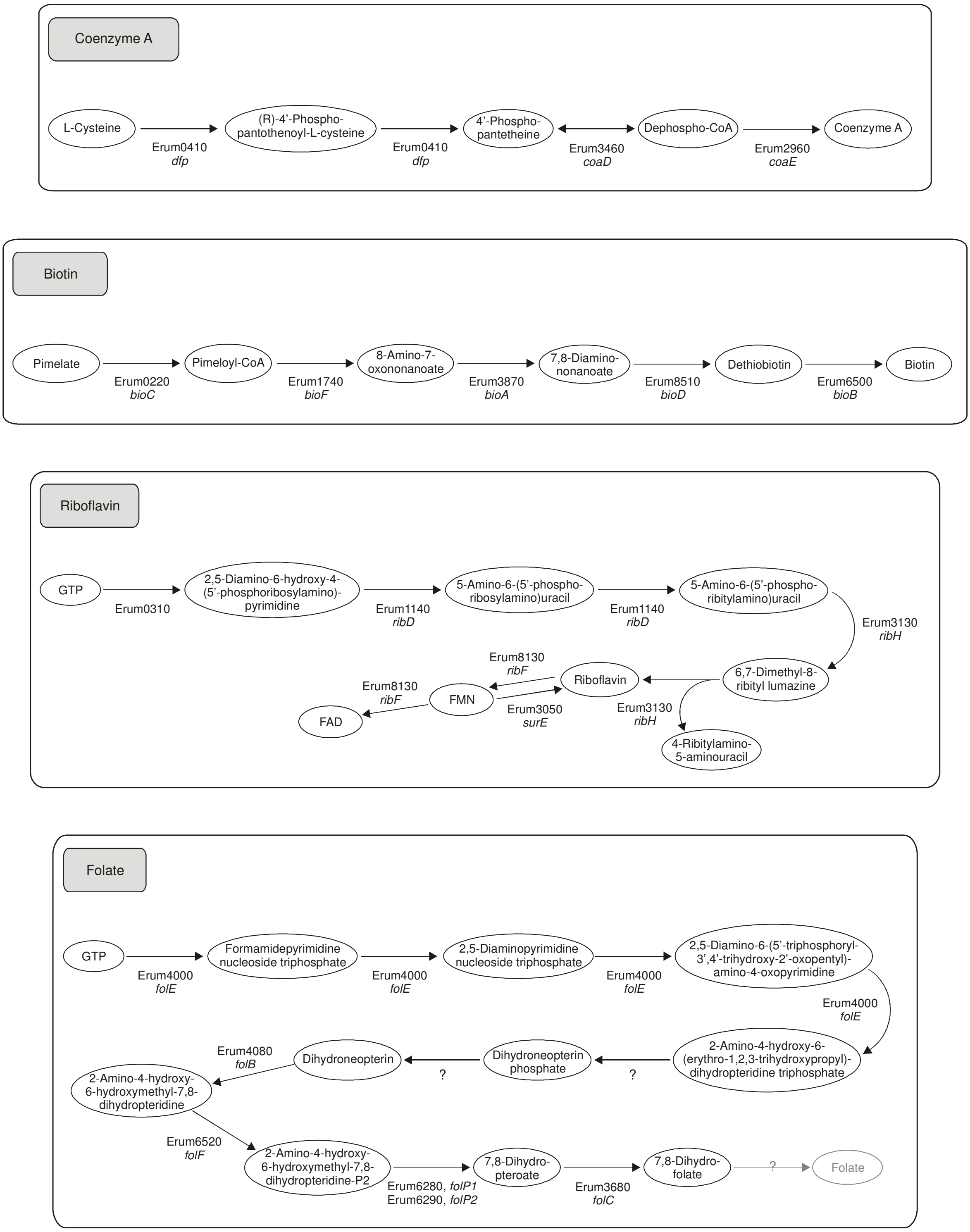

Coenzyme A
L-Cysteine
Erum0410
dfp
(R)-4'-Phospho-pantothenoyl-L-cysteine
Erum0410
dfp
4'-Phospho-pantetheine
Erum3460
coaD
Dephospho-CoA
Erum2960
coaE
Coenzyme A
Biotin
Pimelate
Erum0220
bioC
Pimeloyl-CoA
Erum1740
bioF
8-Amino-7-oxononanoate
Erum3870
bioA
7,8-Diamino-nonanoate
Erum8510
bioD
Dethiobiotin
Erum6500
bioB
Biotin
Riboflavin
GTP
Erum0310
2,5-Diamino-6-hydroxy-4-(5'-phosphoribosylamino)-pyrimidine
Erum1140
ribD
5-Amino-6-(5'-phospho-ribosylamino)uracil
Erum1140
ribD
5-Amino-6-(5'-phospho-ribitylamino)uracil
Erum3130
ribH
6,7-Dimethyl-8-ribityl lumazine
Erum3130
ribH
4-Ribitylamino-5-aminouracil
Riboflavin
Erum8130
ribF
Erum3050
surE
FMN
Erum8130
ribF
FAD
Folate
GTP
Erum4000
folE
Formamidepyrimidine nucleoside triphosphate
Erum4000
folE
2,5-Diaminopyrimidine nucleoside triphosphate
Erum4000
folE
2,5-Diamino-6-(5'-triphosphoryl-3',4'-trihydroxy-2'-oxopentyl)-amino-4-oxopyrimidine
Erum4000
folE
2-Amino-4-hydroxy-6-(erythro-1,2,3-trihydroxypropyl)-dihydropteridine triphosphate
?
Dihydroneopterin phosphate
?
Dihydroneopterin
Erum4080
folB
2-Amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine
Erum6520
folF
2-Amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine-P2
Erum6280, folP1
Erum6290, folP2
7,8-Dihydro-pteroate
Erum3680
folC
7,8-Dihydro-folate
?
Folate

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.

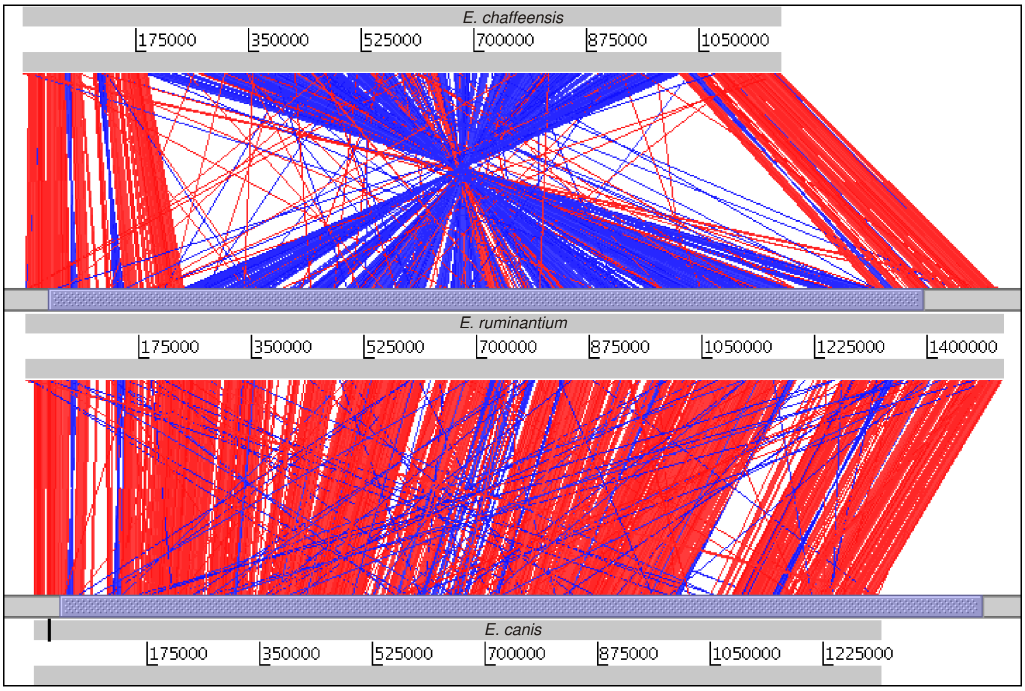


Figure 3.7. Global comparison between *E. ruminantium* (middle), *E. chaffeensis* (top) and *E. canis* (bottom) displayed using ACT.

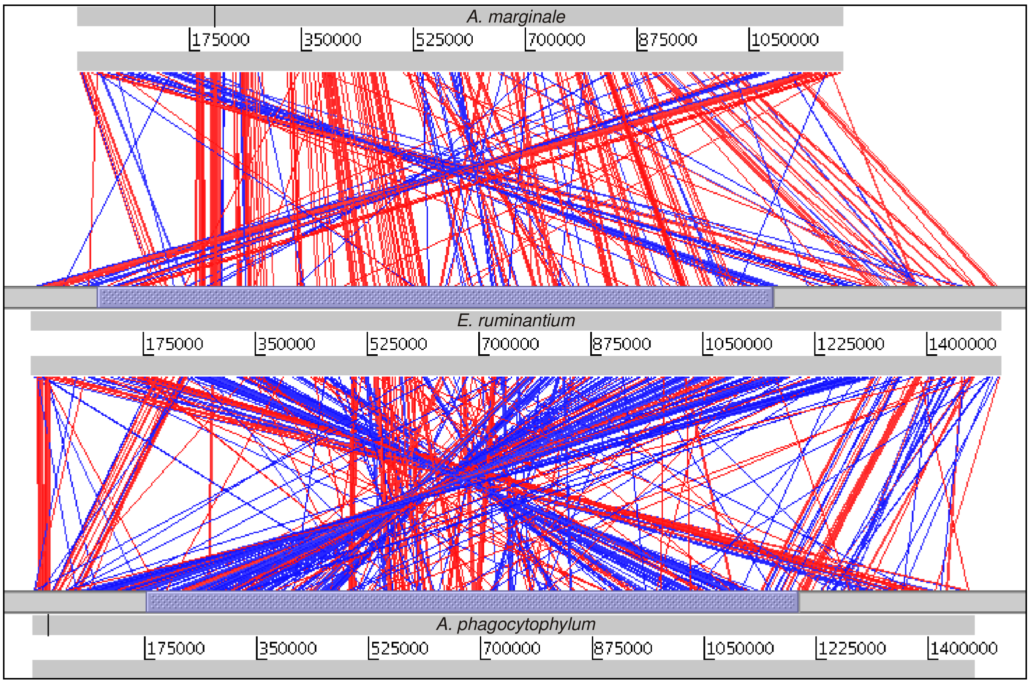


Figure 3.8. Comparison of chromosomal synteny between *E. ruminantium* (middle), *A. marginale* (top) and *A. phagocytophilum* (bottom).

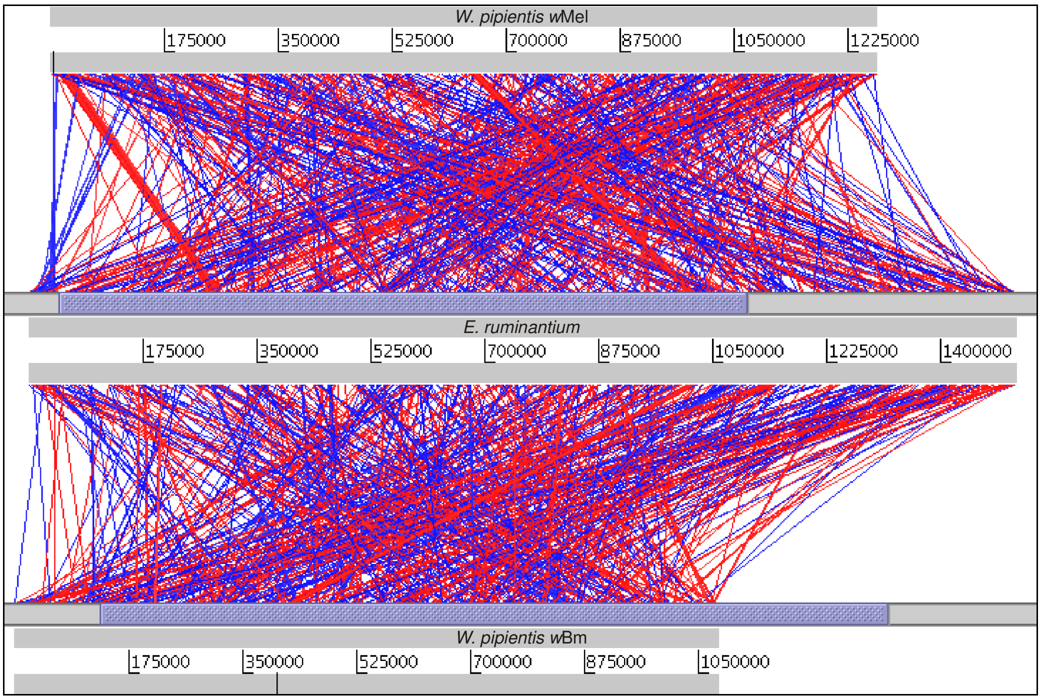


Figure 3.9. Genomic location of the homologous genes in *E. ruminantium* (middle) and the two *Wolbachia* species.

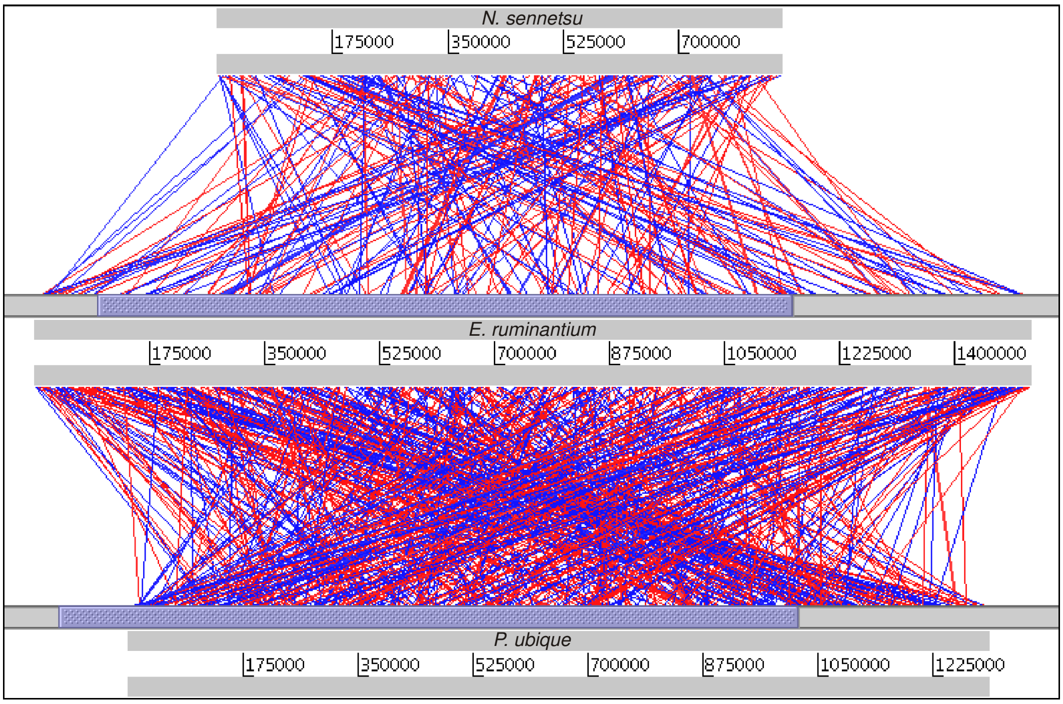


Figure 3.10. *E. ruminantium* gene order compared to *N. sennetsu* (top) and *P. ubique* (bottom).

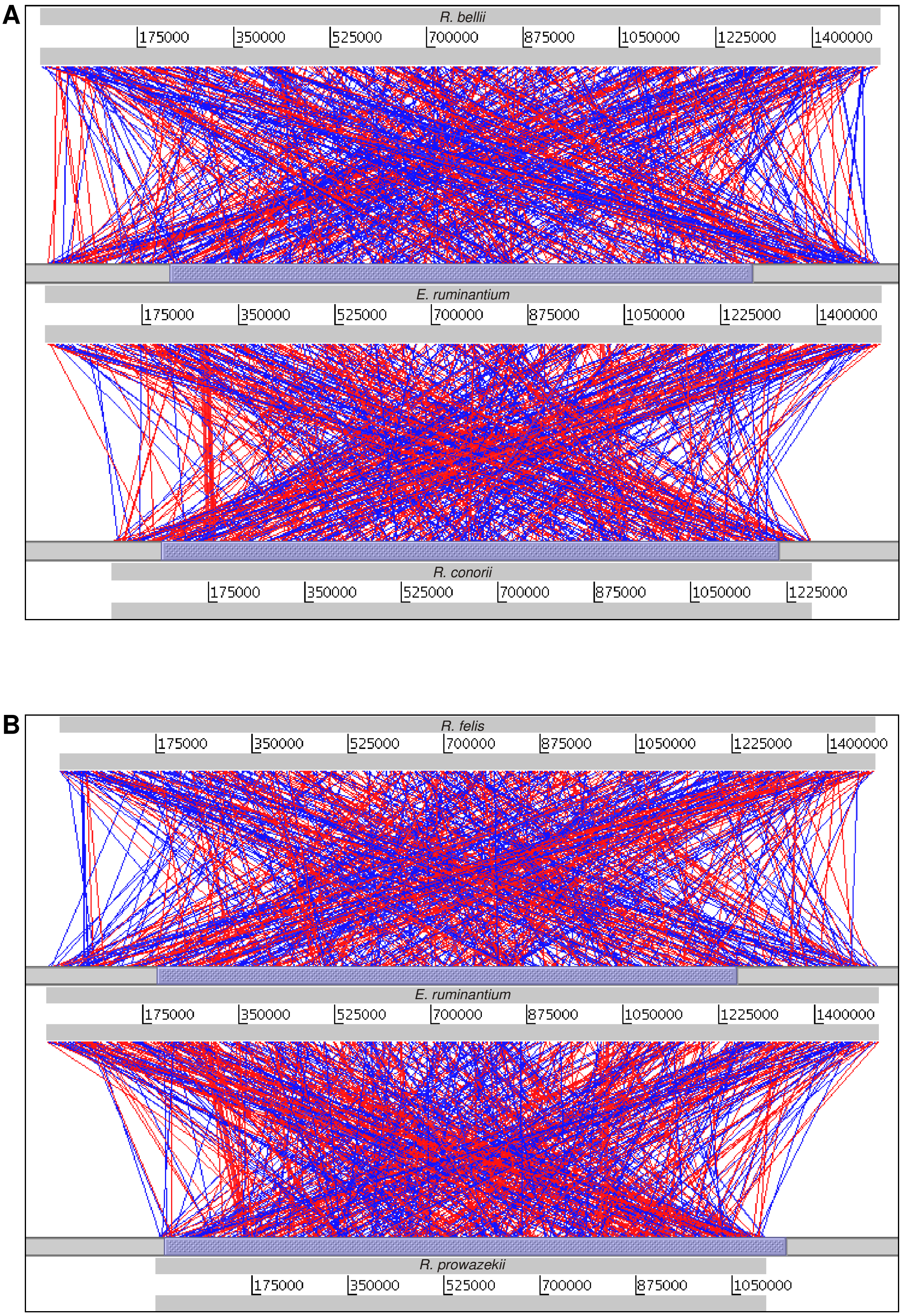


Figure 3.11. A. Comparison of relative positions of conserved genes between *E. ruminantium*, *R. bellii* (top) and *R. conorii* (bottom). **B.** *E. ruminantium* gene order compared to *R. felis* (top) and *R. prowazekii* (bottom).

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	+	+										