CHAPTER 4

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

4.1. INTRODUCTION

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

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

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

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

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

4.2. Materials and Methods

See Appendix B for materials and media components.

4.2.1. Analysis of genomic repeat sequences

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

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

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

4.2.2. Amplification and cloning of variable repeat regions

The following primers were used to amplify the regions containing variable numbers of tandem repeat units: 758_RC1_F and 758_RC1_F (rpt121, Table 4.2); 758_RC2_F and 758_RC2_F (rpt148, Table 4.2); WTHIN440_5F and WTHIN440_5R (rpt18, Table 4.2); and WGAP71walk_1F and WGAP71R (rpt110, Table 4.2). The sequences of these primers can be found in Appendix C1. Template genomic DNA was prepared as described in sub-section 2.2.1.1 and PCR amplifications were conducted using the Platinum® *pfx* DNA polymerase kit (Invitrogen). Each 50 µl reaction contained 25 ng genomic DNA, *pfx* PCR buffer, 0.3 mM dNTPs, 1 mM $MgSO_4$, 0.2 µM of each primer and 1 U *pfx* DNA polymerase. The reaction conditions consisted of one cycle of 5 min at 94°C, 30 cycles of 20 s at 94°C, 30 s at 50°C and 2 min at 68°C, followed by a final incubation of 10 min at 68°C. Amplified products were visualised by electrophoresis on a 1% agarose gel, stained with ethidium bromide. The amplicons were purified with the High Pure PCR Product Purification kit (Roche) and cloned into the pGEM-T Easy vector (Promega) using the protocols provided by the manufacturers. Plasmid DNA was isolated using the High Pure Plasmid Isolation kit (Roche) according to the manufacturer's instructions and digested with *Eco*RI (Roche). The inserts were visualised on 1% agarose gels. At least 20 clones of each region were selected and sequenced with the SP6 and T7 primers (Appendix C4). We sequenced several clones fourfold to show that any observed variation was not an artefact of the sequencing process. Sequencing reaction conditions were as described in sub-section 2.2.1.4.

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

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

4.3. Results and Discussion

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

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

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

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

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

4.3.2. Simple sequence repeats (SSRs)

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

4.3.3. Longer tandem repeats (LTRs)

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

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

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

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

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

4.3.3.1. Tandem repeats in coding regions

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

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

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

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

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

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

4.3.3.2. Repeat regions with variable number of repeat units

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

Of the three 7 bp repeat regions one (rpt18) cannot be translated into ORFs, another (rpt121) can be translated on all three forward frames, and the third (rpt148) has ORFs in all six frames. However, none of the translated ORFs are predicted to be protein-coding. All three 7 bp tandem repeat regions have a higher G+C content than the rest of the genome and exhibit strand asymmetry (one strand contains predominantly either Gs or Cs). Other G+C rich hypervariable sequences have been shown to form secondary structures, which can cause DNA polymerase to pause and may result in the rapid generation of tandem repeats (Weitzmann *et al.*, 1997). The formation of secondary structures thus may explain the variability in the number of these 7 bp repeat units.

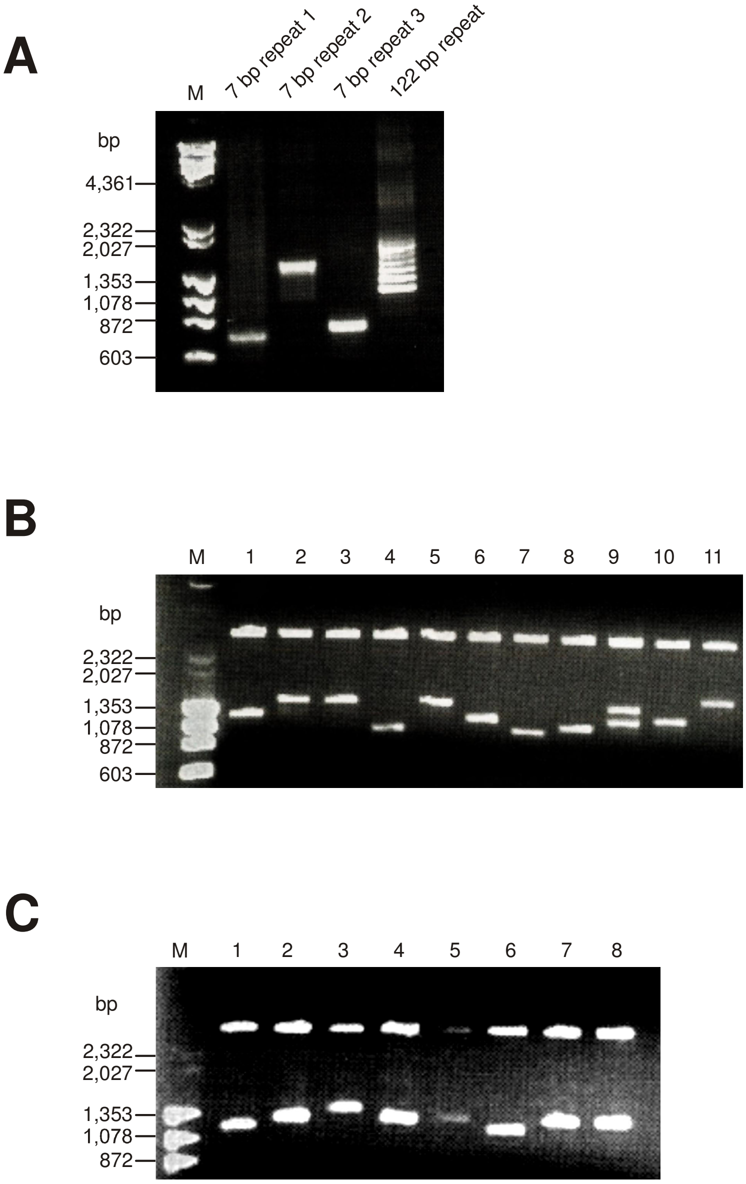

Figure 4.1. Amplification and cloning of variable repeat regions from *E. ruminantium* Welgevonden genomic DNA. **A**. Repeat regions rpt121 (7 bp repeat 1), rpt148 (7 bp repeat 2), rpt18 (7 bp repeat 3), and rpt110 (122 bp repeat) amplified by PCR. **B.** Some of 7 bp repeat 2 (rpt148) clones showing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

4.3.4. Interspersed repetitive DNA

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

4.3.4.1. Homologous recombination between repetitive sequences

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

Another consequence of RecA mediated homologous recombination is gene conversion (Wiuf & Hein, 2000; Chen *et al.*, 2007). This mechanism involves the unidirectional transfer of genetic material from one region to the corresponding place in another paralogous region, which results in homogenisation of the sequences of repeated genes (Petes & Hill, 1988; Lawson *et al.*, 2009; Osada & Innan, 2009). In *E. ruminantium*, gene conversion appears to have limited the divergence between the *rho1* and *rho2* and the *tufA* and *tufB* genes which respectively have 94.0% and 100% identity in overlapping regions. In *Salmonella enterica* serovar Typhimurium such co-evolution of the *tufA* and *tufB* genes has been linked to chromosomal rearrangements (Hughes, 2000b). Both the *rho* and the *tuf* genes are in inverse order on opposite strands of the *E. ruminantium* chromosome and are located on opposite sides of the origin of replication, so recombination between these genes could lead to inversion of the region bounded by the genes. This observation led us to search for such inversions in 12 different *E. ruminantium* isolates, but amplification across the *rho* (Figure 4.2) and *tuf* repeat units with combinations of primers located on either side of the repeat units indicated that the chromosomal arrangement is the same in all the isolates tested. Therefore, although such chromosomal rearrangements may well occur in *E. ruminantium*, the recombinant progeny may not be viable. In fact, although large chromosomal rearrangements are less common within than between species (Hughes 2000a) a high frequency of rearranged genomes has been found in clinical isolates of other pathogenic bacteria such as *Salmonella enterica* serovar Typhi, *Neisseria* spp, *Pseudomonas aeruginosa* and *Bordetella pertussis* (reviewed in Hughes, 2000a). These rearrangements may be favoured as a result of conferring some survival advantage, such as improving the ability of the populations containing them to evade the immune system, however this situation does not seem to have occurred in the case of *E. ruminantium*.

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

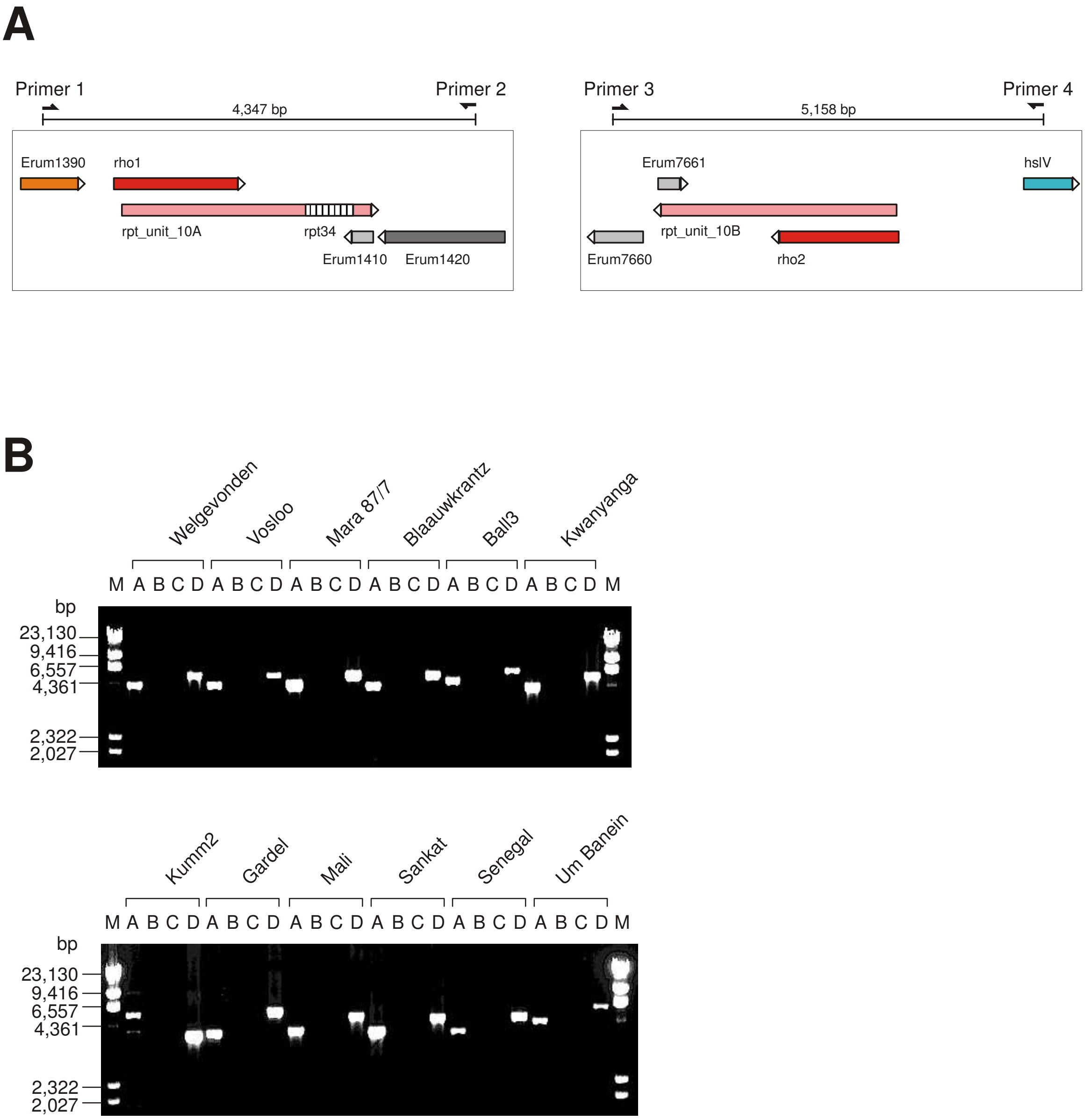

Figure 4.2. PCR amplification across the *rho* repeat regions in *E. ruminantium* isolates.

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

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

4.3.4.2. Duplications appear to generate new genes

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

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

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

In a third example, two genes appear to have been duplicated and fused (Figure 4.3, panel C). The sequence of Erum4120 is 99.1% identical to the 5' end of Erum4140, while the 3' end of Erum4140 has 50.5% identity with the sequence of the following gene Erum4150. It appears likely that Erum4150 was duplicated and the copy mutated until it was 50.5% identical with its parent gene. Subsequently Erum4120 was duplicated and the copy fused with the mutated copy of Erum4120 to generate the new gene Erum4140. The Pfam domains identified in Erum4120 and Erum4150 were also present in Erum4140. Erum4120 is a conserved hypothetical protein and contains a probable transcriptional regulator domain (PF02082), while Erum4150 has similarity to cysteine desulfurase.

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

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

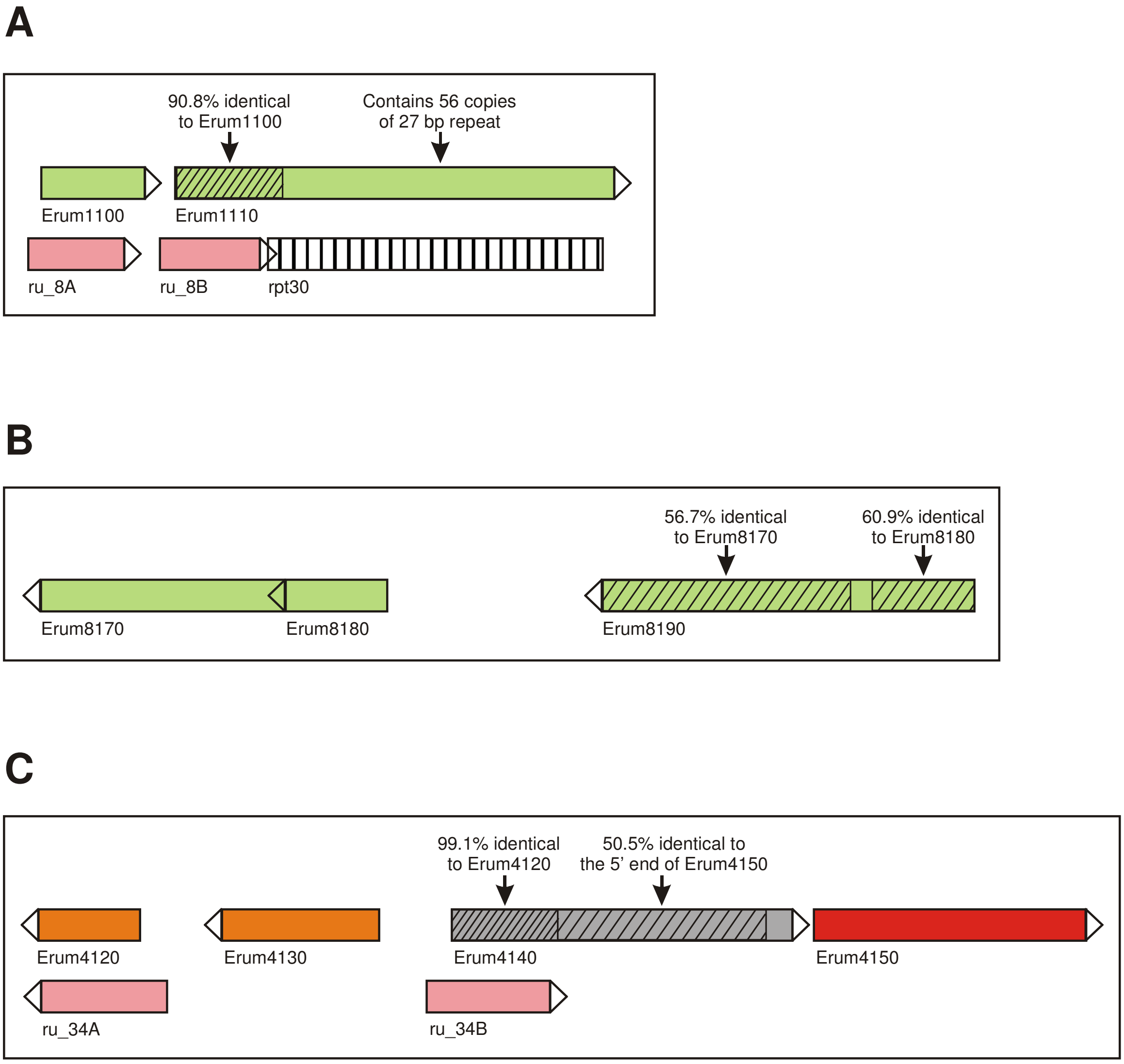
A
90.8% identical to Erum1100
Contains 56 copies of 27 bp repeat
Erum1100
Erum1110
ru_8A
ru_8B
rpt30
B
56.7% identical to Erum8170
60.9% identical to Erum8180
Erum8170
Erum8180
Erum8190
C
99.1% identical to Erum4120
50.5% identical to the 5' end of Erum4150
Erum4120
Erum4130
Erum4140
Erum4150
ru_34A
ru_34B

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

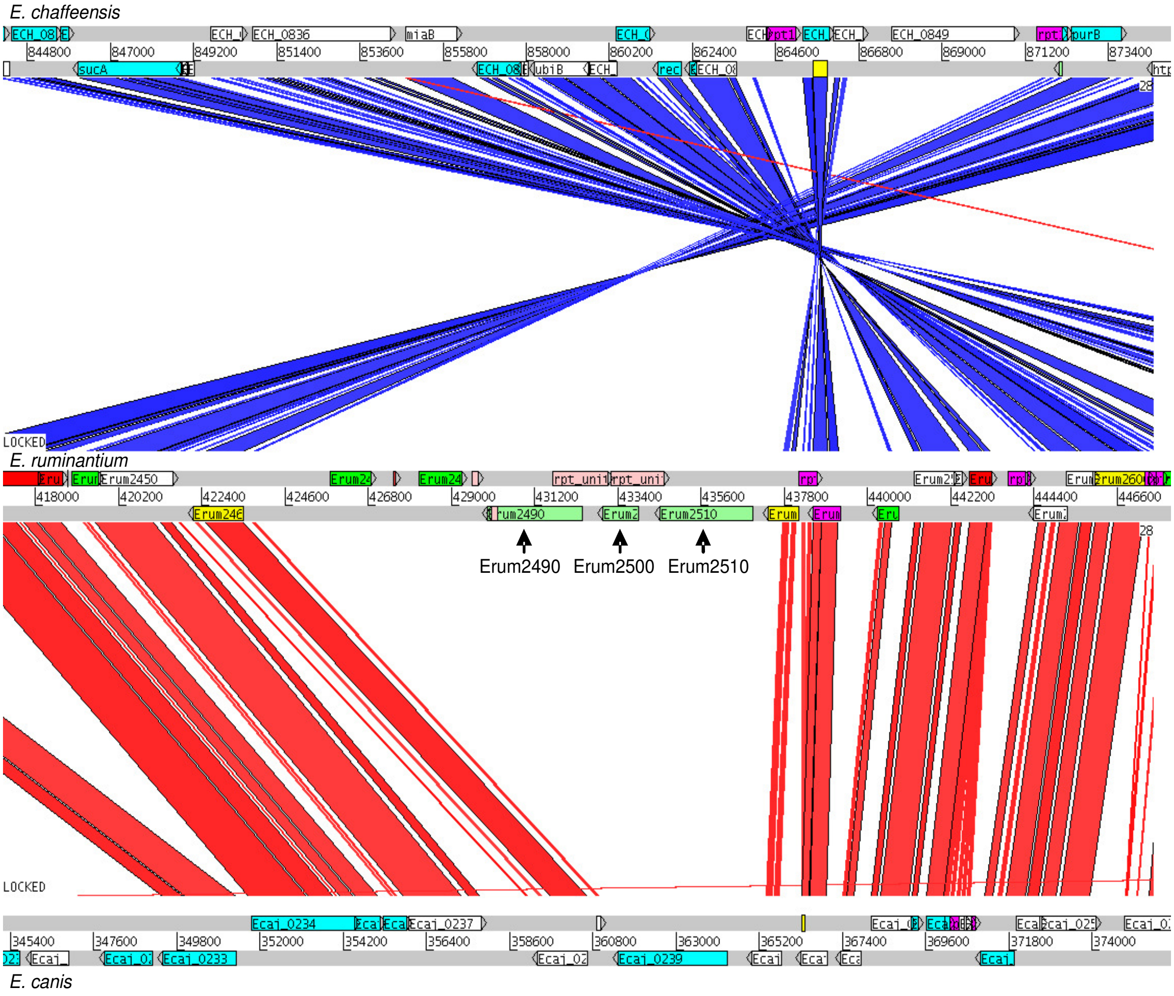

Figure 4.4. Screen capture from ACT of the area around Erum2490, Erum2500 and Erum2510 in *E. ruminantium* (middle), compared to *E. chaffeensis* (top) and *E. canis* (bottom). The grey bars indicate the forward and reverse strands with putative ORFs, while the red and blue lines between the genomes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

4.3.5. Ankyrin repeats

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

4.4. Conclusions

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

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

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

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

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

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

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

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

CHAPTER 5

Selection of possible vaccine candidates

5.1. INTRODUCTION

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

In many instances the traditional methods have failed to generate effective vaccines and yet more modern approaches, such as the development of recombinant subunit or DNA vaccines, have had a limited impact on vaccine production, generating only a few efficacious recombinant vaccines (Grandi, 2003). Examples of commercialised recombinant subunit vaccines include the formulations against *Bordetella pertussis*, hepatitis B virus, *Vibrio cholera*, *Borrelia burgdorferi* and the human papilloma virus (Kaushik & Sehgal, 2008; http://www.fda/gov/). In recent years vaccine development has been revolutionised by the advances in molecular genetics, DNA sequencing and bioinformatics, and the availability of a growing number of complete microbial genome sequences enables the targeting of possible vaccine candidates starting from genomic information, an approach named reverse vaccinology (Rappuoli, 2000).

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

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

5.2. Materials and Methods

See Appendix B for materials and media components.

5.2.1. *In silico* selection strategy

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

5.2.2. Expression of recombinant proteins

5.2.2.1. Directional cloning into the pET vector

Protein expression was performed using the pET102/TOPO® expression system (Invitrogen). Sequence specific primers (Appendix C3) were designed for each of the selected ORFs to facilitate directional cloning into the pET vector. In the case of ORFs having signal peptide coding sequences the 5' primers were designed so as to omit the signal sequences. ORFs larger than 2,000 bp were divided into smaller sub fragments and we also made sure that primer sequences did not overlap large tandem repeat sequences. The ORFs were amplified with *Pfu* polymerase (Promega), a proofreading DNA polymerase that produces blunt ended PCR products. Each 50 µl reaction contained 25 ng *E. ruminantium* (Welgevonden) genomic DNA, 1.25 U *Pfu* polymerase, 0.2 µM of each primer, 0.2 mM dNTPs, and 1x reaction buffer (containing 2 mM Mg^{2+}). Amplification was carried out on a GeneAmp® PCR System 9700 (Perkin-Elmer Applied Biosystems) under the following conditions: one cycle at 95°C for 2 min, followed by 35 cycles of denaturation (95°C for 30 s), annealing (50°C for 30 s), and extension (72°C for 3 min), with a final extension at 72°C for 7 min. The amplicons were purified with the MSB® Spin PCRapace kit (Invitek) and cloned into the TOPO® pET vector following the manufacturer's protocols. The plasmid constructs were transformed into TOP10 competent *E. coli* cells by electroporation using the Gene pulser™ II (Bio-Rad) as described in the manufacturer's manual. The cells were plated on LB agar plates containing 50 µg/ml ampicillin and incubated overnight at 37°C. Recombinant clones were picked and grown overnight at 37°C in LB broth containing ampicillin (50 µg/ml). The plasmid DNA was purified using the Invisorb® Spin Plasmid Mini *Two* kit (Invitek) and inserts were detected by PCR followed by 1% agarose gel electrophoresis of the amplicons. The reaction mix contained 0.5 µl of plasmid DNA, 0.13 U TaKaRa Ex Taq™ (TaKaRa Bio Inc.), 0.2 mM dNTPs, and 0.25 µM of each of the pET vector specific primers, TrxFus forward and T7 reverse (Appendix C4). The reaction conditions were: one cycle at 94°C for 5 min, 35 cycles of 94°C for 30 s, 50°C for 30 s and 72°C for 3 min, and a final extension at 72°C for 7 min. Clones containing inserts of the correct size were sequenced, using the TrxFus forward and T7 reverse primers, to verify the orientation and sequences of inserts and to ensure that the His-tag was in-frame.

5.2.2.2. Expression and purification of recombinant proteins

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

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

5.2.2.3. Western blot analysis

Expressed proteins were analysed by Anti-His$_6$ Western blot analysis using standard procedures. The purified proteins were separated on Criterion™ XT precast gels (4-12% gradient, Bio-Rad) at 100 V for approximately 2 h and transferred to PVDF membranes (Millipore Corporation) with a semi-dry blotter (Semi-phor TE70, Hoefer Scientific Instruments) at 110 mA for 90 min. After incubation in blocking buffer (1x PBS, 1% BSA) for 1 h, the blots were incubated overnight at room temperature in the presence of Anti-His$_6$ antibodies (75 ng/100 ml, Roche) and the following day they were exposed to conjugate [1/20,000 dilution, horseradish peroxide-goat-anti-mouse IgG (Zymed)] for 1 h at room temperature. The membranes were washed three times with wash buffer (1x PBS, 0.05% Tween-20) for 5 min after each incubation step. Finally the recombinant His-tagged protein bands were visualised using SuperSignal®West Pico Chemiluminescent substrate (Pierce) and X-ray film (Roche).

5.2.3. Immunological assays

5.2.3.1. Lymphocyte proliferation assays

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

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

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

5.2.3.2. IFN-γ ELISpot assays

IFN-γ expression was measured by enzyme-linked immunospot (ELISpot) assays in 96-well plates. MAIPS 4510 Multiscreen™-IP filtration plates (Millipore) were coated overnight with mouse anti-bovine IFN-γ mAb CC302 (1 µg/ml) (Celtic Molecular Diagnostics) at 4°C, and washed three times with unsupplemented RPMI-1640. The coated wells were blocked with RPMI-1640 supplemented with 10% FCS for 2 h at 37°C. Freshly isolated PBMCs (4 x 10^6/ml) were added to the wells and stimulated with the recombinant proteins (1 µg/ml) and incubated for 20 h at 37°C in a humidified atmosphere with 5% CO_2. Positive (ConA) and negative (no antigen) controls were included, as already described for the proliferation assays. The plates were washed three times with 0.05% dH_2O-Tween, three times with 0.05% PBS-Tween (PBS-T) and incubated with rabbit anti-bovine IFN-γ anti-serum (Immonodiagnostik) diluted 1/1,500 in PBS-T/1% BSA for 1 h at room temperature. Subsequently the plates were washed four times with 0.05% PBS-T, followed by incubation for 1 h at room temperature with anti-rabbit IgG alkaline phosphatase conjugate (Sigma) diluted 1/2,000 in PBS-T/1% BSA. After six washes with 0.05% PBS-T, 50 µl of substrate solution (Sigma Fast BCIP/NBT substrate tablets) were added and the plates were incubated in the dark for 15 min. The plates were washed for 2 min under running water and dried overnight. Spot forming cells (SFCs) were counted using an automated ELISpot reader (Zeiss KS ELISPOT Compact 4.5). The number of SFCs produced after stimulation of immune PBMCs with the recombinant proteins was compared to the number of SFCs produced after stimulation of naïve PBMCs with the recombinant proteins. ELISpot samples with 4x the number of spots/million cells compared to the naïve cells were considered positive.

5.2.4. Vaccine trials in sheep

5.2.4.1. Challenge material

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

5.2.4.2. DNA immunisation

5.2.4.2.1. Cloning of ORFs into pCMViUBs

The ORFs were amplified using specific primers (Appendix C3) containing restriction enzyme sites to facilitate directional cloning into the pCMViUBs vector (Sykes & Johnston, 1999). We searched the sequences of the ORFs for internal restriction sites using the Staden package program Spin (Staden *et al*., 2000). Of the available recognition sites incorporated in the vector's cloning site, the cutting sites of the endonucleases *Bam*HI and *Sal*I were not present in any of the ORF sequences and were therefore integrated into the primer sequences. PCR amplifications were performed in 100 µl reaction mixtures containing: 50 ng genomic *E. ruminantium* (Welgevonden) DNA, 0.2 mM dNTPs, 0.25 µM of each primer and 0.5 U TaKaRa Ex Taq™ (TaKaRa Bio Inc.) in 1x reaction buffer. The samples were denatured for 2 min at 95°C, followed by 35 cycles of 95°C for 30 s, 50°C for 30 s, and 72°C for 3 min; this was followed by a final extension at 72°C for 10 min. The PCR products were purified with the MSB® Spin PCRapace kit (Invitek) and cloned into the pGEM®-T Easy vector system (Promega) using the protocols provided by the manufacturers. Recombinant cells were grown overnight at 37°C in LB broth containing 50 µg/ml ampicillin. The plasmid DNA was purified using the Invisorb® Spin Plasmid Mini *Two* kit (Invitek), digested with *Eco*RI (Roche) and inserts were visualised on 1% agarose gels. Clones containing fragments of the expected size were sequenced with the SP6 and T7 primers (Appendix C4). Plasmids containing the correct insert sequence were digested with *Bam*HI and *Sal*I, while the pCMViUBs vector was digested with the same enzymes and dephosphorylated using shrimp alkaline phosphatase (Promega). The ORF inserts and prepared pCMViUBs vector were purified from agarose gels using TaKaRa recochips (TaKaRa Bio Inc.). The inserts were ligated into the linearised dephosphorylated vector using 1 U T4 DNA ligase (Promega). The ligated products were electroporated into TOP10 *E. coli* cells (Invitrogen) using the Gene pulser™ II (Bio-Rad), plated onto LB agar plates containing ampicillin (50 µg/ml) and incubated overnight at 37°C. Positive clones were screened by PCR and sequenced with the vector specific primers, IECO and CMV991 (Appendix C4), to determine whether the correct ORF sequence was present and in-frame.

5.2.4.2.2. Large scale DNA preparation

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

5.2.4.2.3. DNA immunisation of sheep

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

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

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

5.2.4.3. DNA prime–recombinant protein boost immunisation strategy

5.2.4.3.1. Large scale preparation of recombinant proteins

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

5.2.4.3.2. Immunisation of sheep

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

Table 5.1. The immunisation strategy for the animal trial.

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

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

† ORF cocktail 2: Erum5400, Erum8050, Erum5270

5.3. Results and Discussion

5.3.1. *In silico* selection of possible vaccine candidates

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

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

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

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

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

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

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

5.3.2. Expression of recombinant proteins

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

5.3.3. Physical characteristics of recombinant proteins

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

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

5.3.4. Recombinant proteins inducing specific Th1 cellular immune responses

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

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

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

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

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

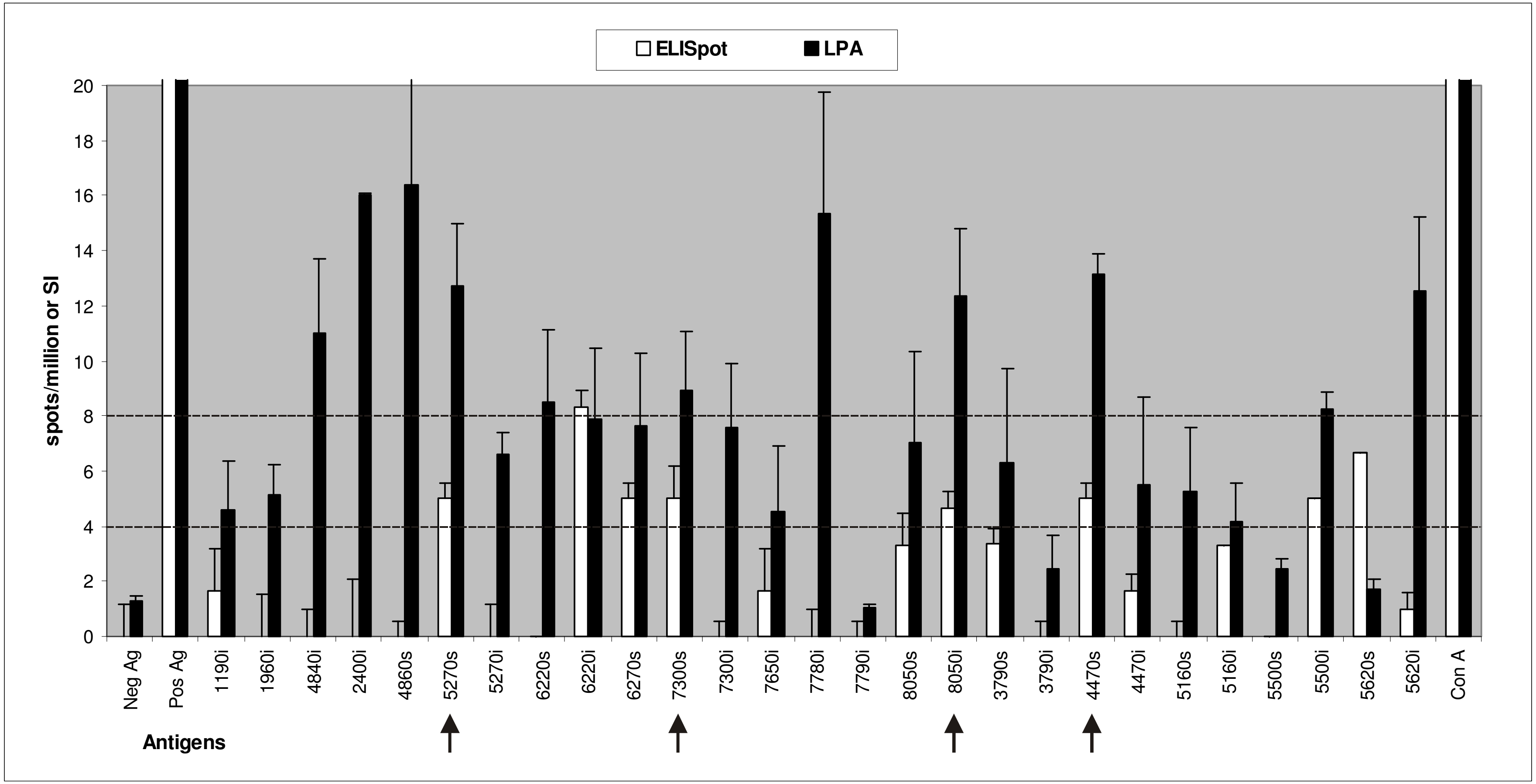

Figure 5.1. ELISpot and lymphocyte proliferation assays (LPA) of PBMCs stimulated with recombinant proteins (plate 1). The **s** and **i** after protein numbers indicate the soluble or insoluble fractions of the proteins. White bars represent the IFN-γ production as spots/million cells, while black bars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

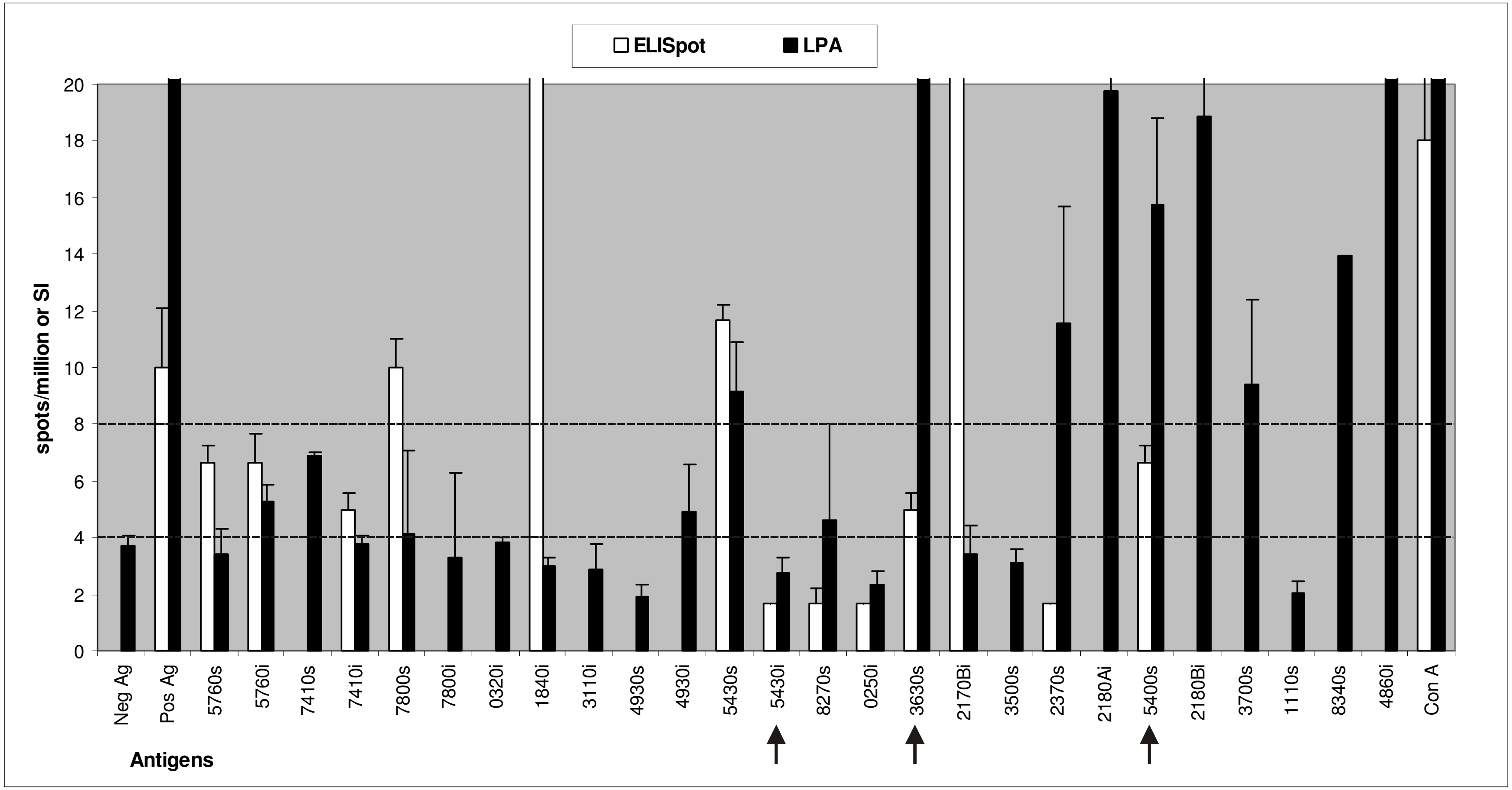

Figure 5.2. ELISpot and lymphocyte proliferation assays (LPA) of PBMCs stimulated with recombinant proteins (plate 2). The **s** and **i** after protein numbers indicate the soluble or insoluble fractions of the proteins. White bars represent the IFN-γ production as spots/million cells, while black bars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

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

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

5.3.5. Vaccine trials in sheep

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

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

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

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

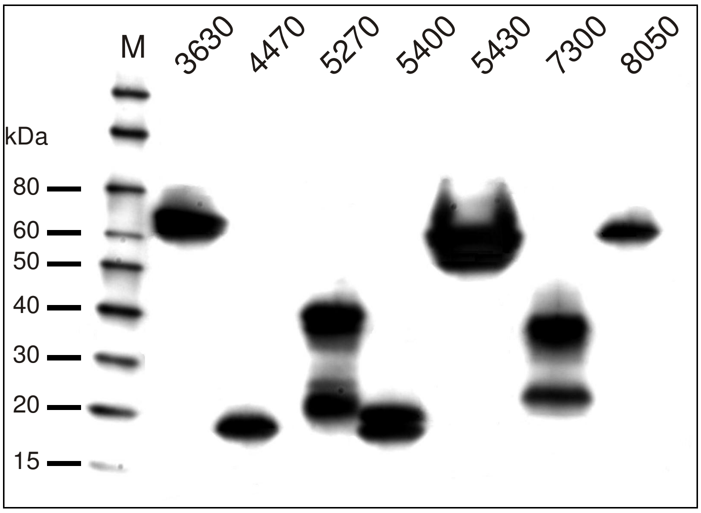

Figure 5.3. Anti-His$_6$ Western blot of the seven selected ORFs expressed in *E. coli*. Lane M = BenchMark™ His-tagged Protein Standard (Invitrogen)

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

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

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

It has been shown that recombinant protein boosting after primary DNA immunisation can enhance protection against pathogens such as *Mycobacterium tuberculosis* (Wang *et al*., 2004a) and *Leishmania infantum* (Rafati *et al*., 2006). In experiments using the *E. ruminantium* 1H12 ORFs both the recombinant DNA-only immunisation, as well as the recombinant DNA priming followed by recombinant protein boosting, provided 100% protection in laboratory conditions (Pretorius *et al*., 2007; 2008). However only lymphocytes isolated from animals which received a protein boost showed specific proliferation and increased IFN-γ expression when exposed to the recombinant proteins (Pretorius *et al*., 2008). Others have also found that boosting with recombinant protein improved lymphocyte proliferation and increased IFN-γ production (Wang *et al*., 2004b; Rafati *et al*., 2006). In another experiment, using the *cpg*1 gene (Erum2510), protein boosting improved the protection against *E. ruminantium* challenge (Pretorius *et al*., 2010). In the current study, however, protein boosting did not confer any protection. The one immunised animal which survived without treatment was in the Experimental 2 group, which had received cocktail 2 by DNA-only immunisation, while no animals survived without treatment in the Experimental 4 group, which had also received cocktail 2, in this case via the DNA prime–recombinant protein boost regimen. It is possible that the immunological mechanism responsible for protection is different for individual genes, for instance, it was suggested that *cpg*1 may activate a humoral response (Pretorius *et al.*, 2010). From the vaccine development viewpoint this is very disappointing since it complicates the practical experimental issues enormously.

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

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

5.4. CONCLUSIONS

Bioinformatic tools were used to identify possible vaccine candidates from the annotated *E. ruminantium* genome sequence. The protective properties of seven ORFs, which induced two different cellular immune responses *in vitro*, were tested in sheep. Only 20% survival was obtained in sheep immunised three times with a DNA formulation consisting of three ORFs; all the other animals succumbed to lethal challenge. The fact that the levels of PBMC proliferation and IFN-γ production did not correlate with each other, nor with the levels of protection, suggests that the current methods being used to select vaccine candidates are just not reliable. In particular it appears that IFN-γ expression alone is not an indicator of protection. We would therefore suggest that other cytokines will have to be included in future immunological studies of the mechanism of protection against *E. ruminantium* to define in detail what constitutes a protective immune response against this organism. Although reverse vaccinology has been applied successfully in a number of studies the approach was not successful in this study and it still remains a challenge to identify suitable *E. ruminantium* vaccine candidates for further investigation.

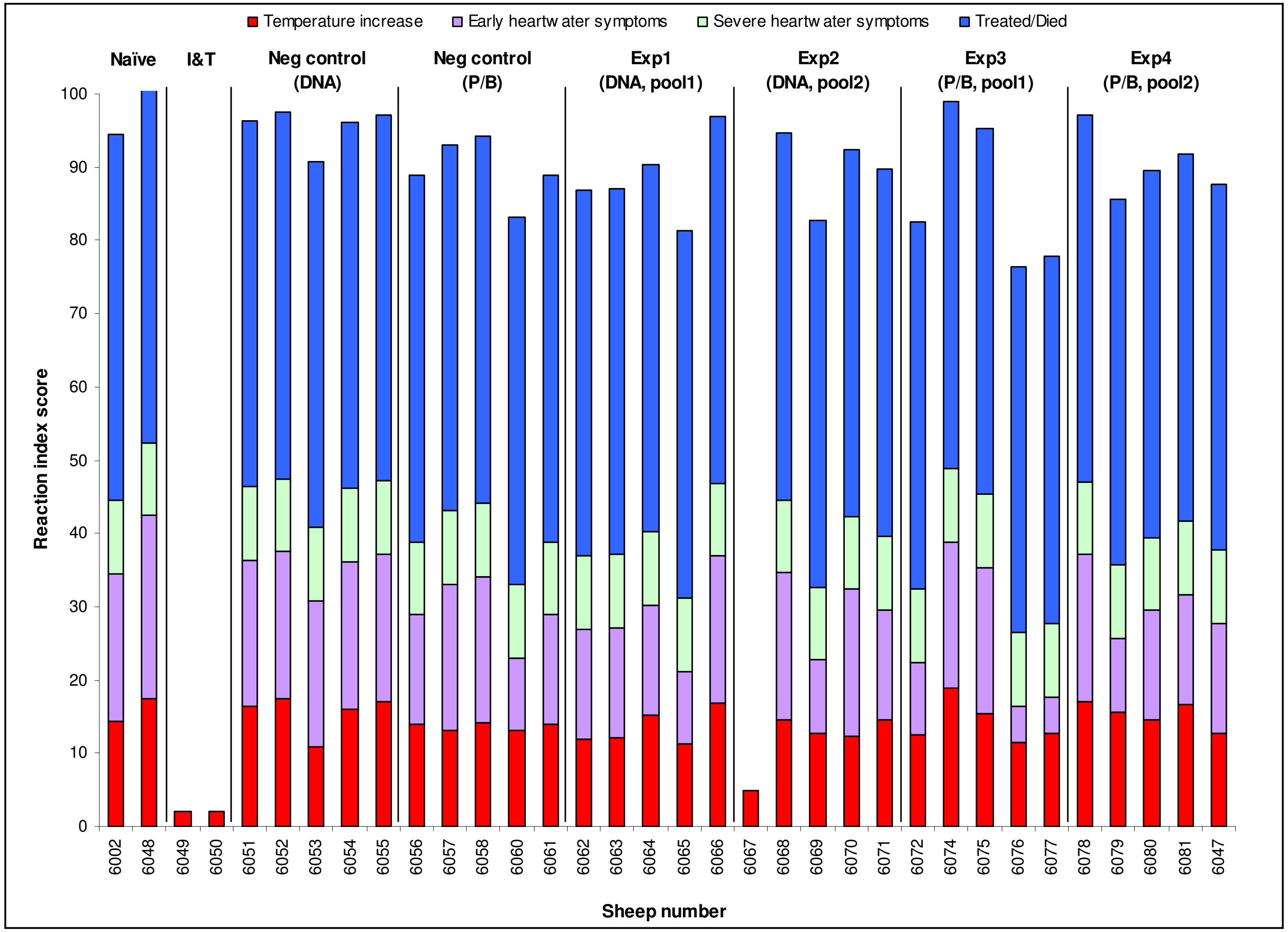

Figure 5.4. Reaction index of sheep. Red blocks represent the total temperature reaction score, while purple indicates early heartwater symptoms, green severe heartwater symptoms, and blue that the animal was treated or euthanased, or died.

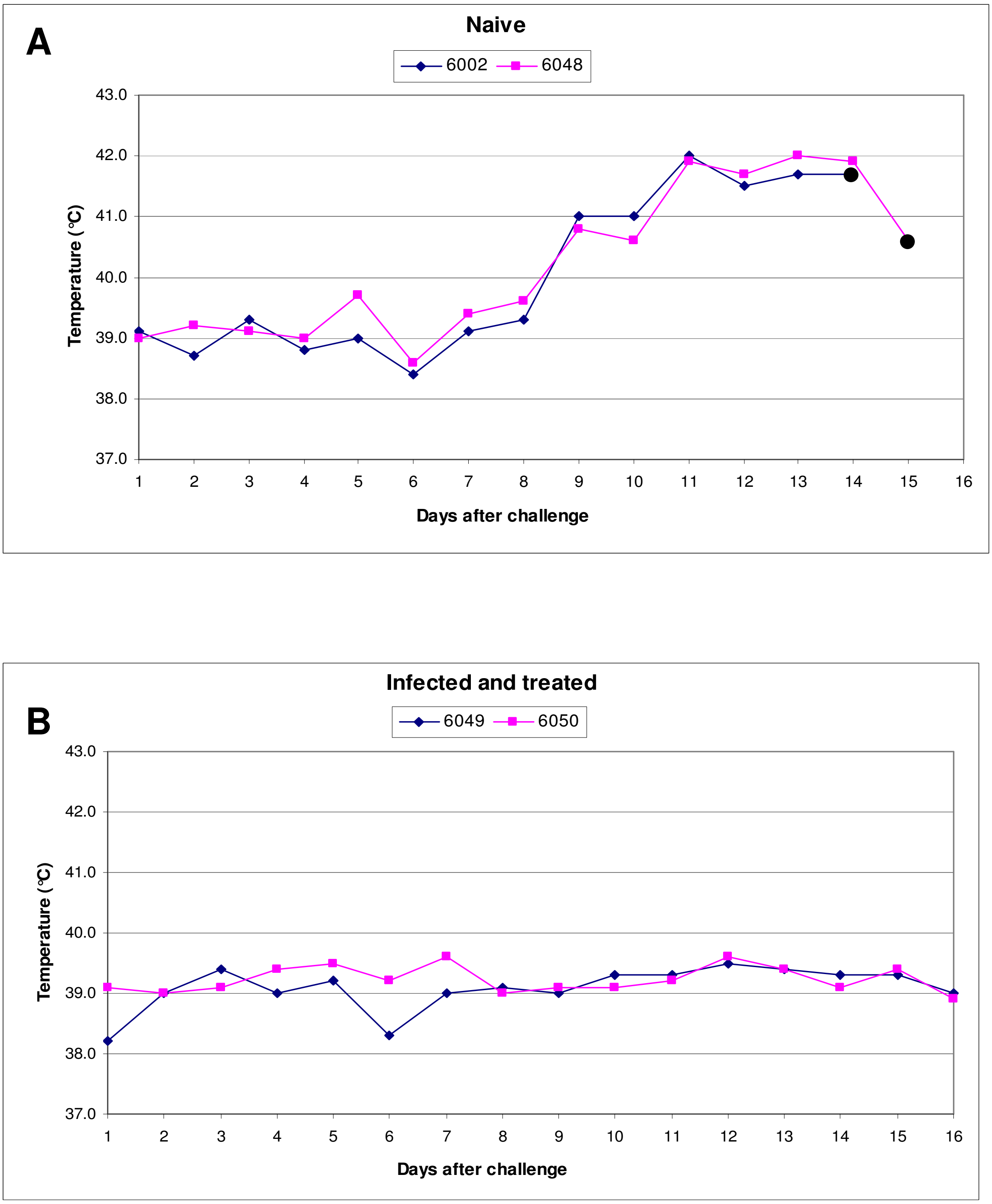

Figure 5.5. Daily post-challenge temperatures of the challenge control group (**A**) and the infected and treated group (**B**). Black dots indicate the day on which the animal died or when it was euthanased. Although both sheep in the infected and treated group survived, temperature measurements are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

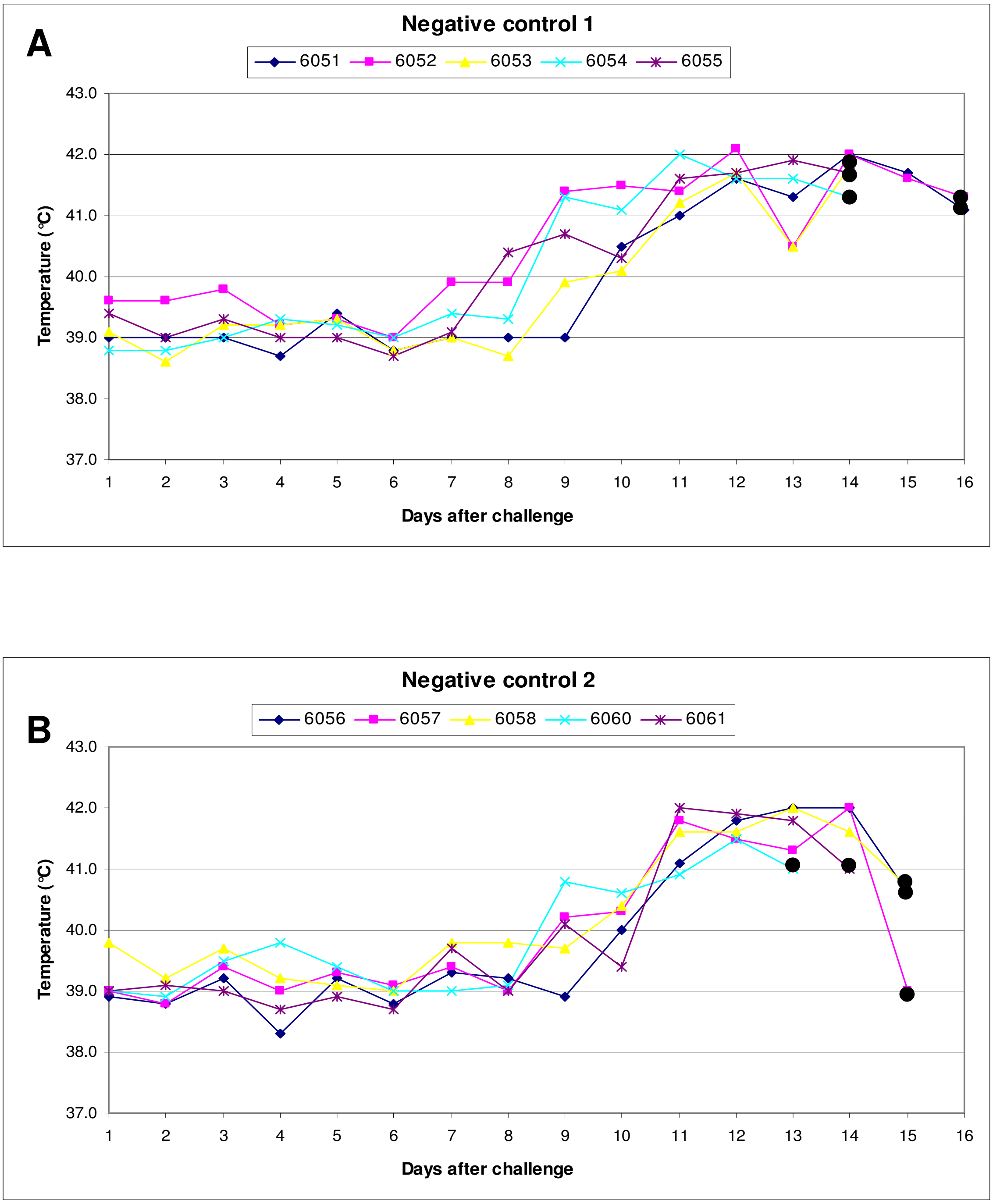

Figure 5.6. Daily post-challenge temperatures of the negative control groups. **A:** Sheep inoculated 3x with empty pCMViUBs vector. **B:** Sheep inoculated twice with empty pCMViUBs vector, followed by a recombinant β-galactosidase protein boost. Black dots indicate the day on which the animal died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

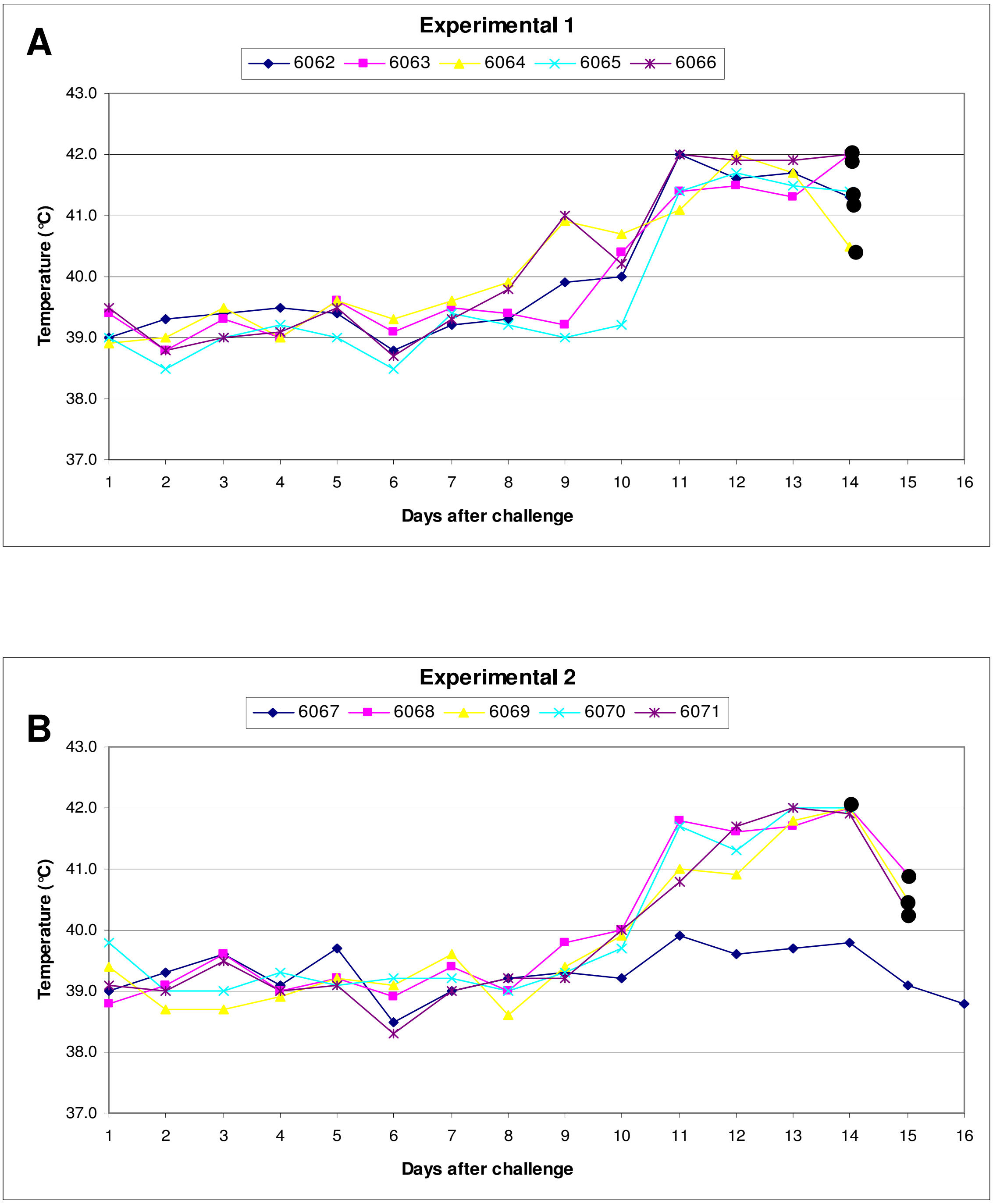

Figure 5.7. Daily post-challenge temperatures of sheep inoculated 3x with ORF cocktail 1 (**A**) or ORF cocktail 2 (**B**) DNA. Black dots indicate the day on which the animal died or when it was euthanased. Temperature measurements of the sheep that survived are shown only until day 16.

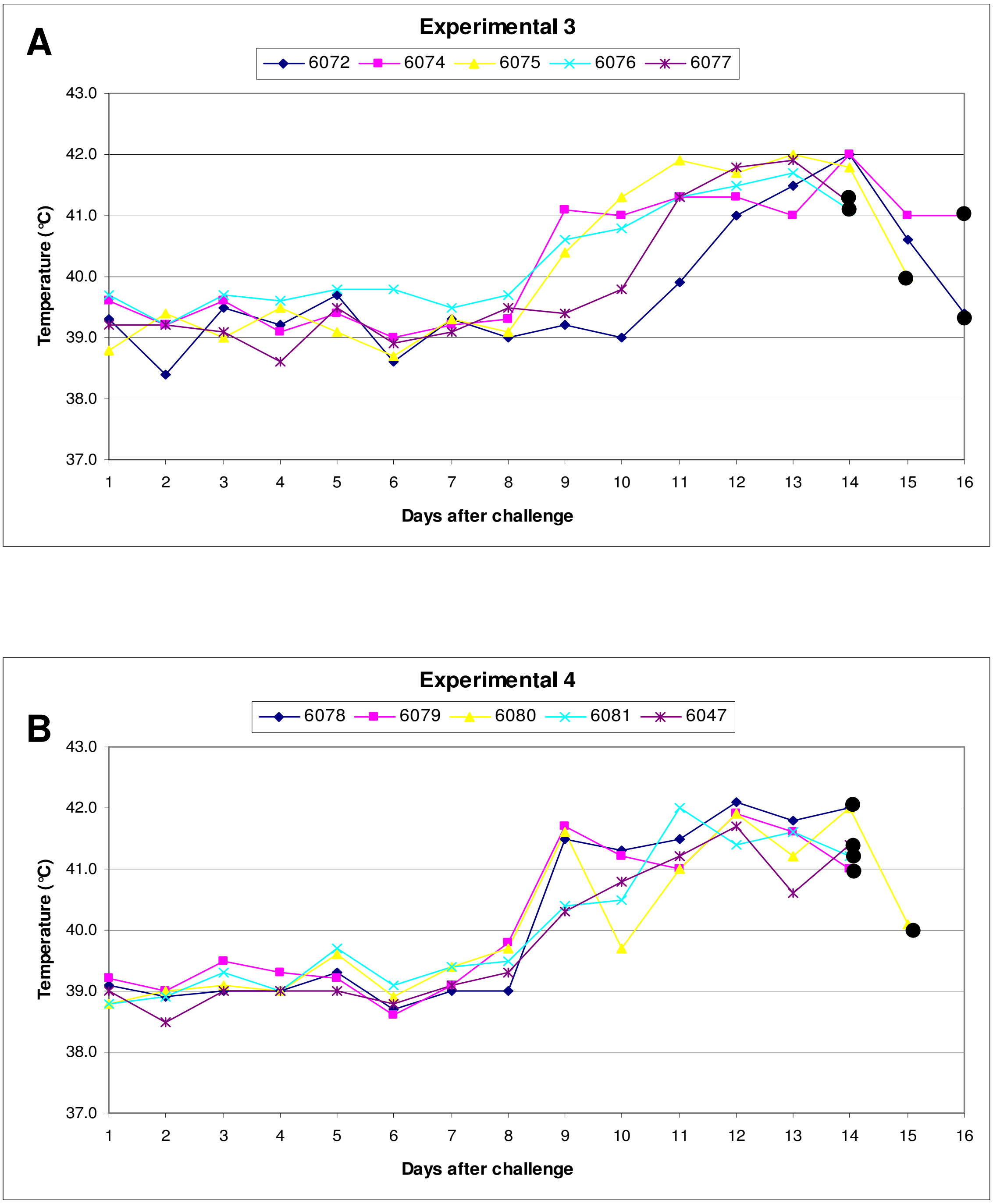

Figure 5.8. Daily post-challenge temperatures of the prime–boost vaccinated groups. **A:** Sheep immunised twice with ORF cocktail 1 DNA followed by an ORF cocktail 1 recombinant protein boost. **B:** Sheep immunised twice with ORF cocktail 2 DNA followed by an ORF cocktail 2 recombinant protein boost.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

CHAPTER 6

Concluding discussion

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

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

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

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

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

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

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

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

Host immune responses to *Anaplasma* infection have been studied by way of expression profiling. For example, it was found that *A. phagocytophilum* infection in sheep modifies host gene expression and immune responses by activating the inflammatory and innate immune pathways and also impairs adaptive immunity (Galindo *et al.*, 2008). Zivkovic and colleagues determined the effect of *A. marginale* infection on gene expression in the salivary glands of *Rhipicephalus microplus* and discovered genes encoding for putative proteins that are probably required by *A. marginale* for infection and multiplication in ticks (Zivkovic *et al.*, 2010). The genome sequences of several vector and host species have also been completed or are in progress. A draft assembly of the tick *Ixodes scapularis*, vector for the Lyme disease spirochete *Borrelia burgdorferi*, is available and sequencing of the *A. marginale* vector, *Rhipicephalus microplus*, is in progress. Also available are the genomes of the bovine (Bovine Genome Sequencing and Analysis Consortium, 2009) and sheep hosts. The combination of pathogen, vector and host sequence data present new prospects to characterise the inherent structural differences that affect host–pathogen interactions, and to study metabolic and immunologic pathways implicated in resistance to infection and disease pathology (Zarlenga & Gasbarre, 2009). Investigation of the host–pathogen–vector interactions via transcriptome analyses may also bring us closer to dual-action vaccines for the control of both pathogen transmission and tick infestation (De la Fuente *et al.*, 2010).

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

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

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

Finally, whole genome sequencing has become a standard method for studying living organisms and, since the first complete genome of a free-living organism, *Haemophilus influenzae*, was obtained in 1995, the number of genome sequences in public databases has grown exponentially. To date 1,181 complete bacterial sequences are available in GenBank and more than 3,300 are being sequenced (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi, as of July 2010). The availability of the *E. ruminantium* genome sequence, the first complete genome of a free-living organism to be sequenced and annotated in Africa, will greatly facilitate novel approaches to the study of the organism and its interaction with its hosts. The data derived from this study are vital resources in the search for an efficacious, cost-effective and practical vaccine against heartwater.