Analysis of the *Helicobacter mustelae* Surface Ring (*hsr*) Locus

A thesis presented in partial fulfilment of
the requirement for the degree of
Master of Science in Biological Sciences
at Massey University, Palmerston North,
New Zealand

Natasha Talei Forester

2001

ABSTRACT

The DNA upstream of the gene encoding the *Helicobacter mustelae* surface ring (Hsr) protein of the ferret gastric pathogen *Helicobacter mustelae* was shown to contain several uncharacterised *hsr*-like repeat sequences in a total of 12 kb of *hsr*-related DNA, including the *hsr* gene (the *hsr* locus). The primary objective of this study was to analyse the *hsr* locus of *H. mustelae* strain 4298, in particular, to determine the extent and possible function of the *hsr*-related DNA repeat sequences.

H. mustelae was isolated from the stomachs of six New Zealand ferrets. This study represents the first successful isolation of *H. mustelae* from ferret stomachs from at least two geographically distinct locations in New Zealand. The Hsr proteins of the different *H. mustelae* strains exhibited variability in protein size and reactivity to anti-Hsr antisera. The DNA sequence of the strain 4298 15-kb *hsr* locus was completed and analysed for DNA repeats. There were 343 perfectly repeated sequences 12 – 741 bp in length, with up to 11 copies of each. Within the *hsr* gene of strain 4298, a 2.4 kb repeat region, the variable repeat region, was defined. The flanking *hsr*-related sequences were equally distributed and mostly (96%) inverted with respect to the *hsr* gene. DNA sequence alignments of nine different *H. mustelae* strains, showed a high level of sequence variation in the variable repeat region of the *hsr* gene, in contrast to the central and β domains. Alignment of sequenced DNA from the variable repeat region of different strains identified conserved-variable-conserved blocks (CVCs) of sequence, which may facilitate a recombination-based antigenic variation mechanism. Approximately 7 kb upstream and 3 kb downstream *hsr*-related flanking sequence may serve as a reservoir for sequence variation of the *hsr* gene. The searches for repeat elements have facilitated the identification of potential DNA regulatory elements involved in the abundant production of the Hsr protein.

The HSRL also contained an unrelated open reading frame, encoding Orf2, which had significant identity with LolA , a periplasmic lipoprotein carrier protein, but containing an N-terminal extension of 14 charged and polar amino acids. Insertional inactivation of *orf2* had no detectable effect on Hsr expression in the Hsr^+ strain 4298.

ACKNOWLEDGEMENTS

To my supervisor, **Dr Paul O'Toole**. First, thank you for giving me the opportunity to work in your lab. I have enjoyed the experience. Thanks also for endeavouring to provide various resources for the lab to keep us all happy/quiet. Many thanks for the scientific guidance and taking time to read and provide helpful suggestions during the writing of this thesis. I apologise for the all the Tash-induced migraines received over the past few years.

Thanks also to **the Institute of Molecular BioSciences/ Department of Microbiology and Genetics** for allowing me to do this study part time while working. Thanks also to **Dr Kathryn Stowell** for cheerfully providing helpful advice on a number of occasions.

I would like to thank **Dr Kathy Parton** (IVABS, Massey University) for sourcing ferret samples, without which, I would not have been able to complete this work. Many thanks also to **Mr Terry Hynes** (District manager, Agriquality New Zealand) for the taking the time to collect and send ferret samples.

I am grateful to all Helipad members over the past four years (**Grover, Amanda, Kirsty, James, Mick, Basil, Michael (x2), Millis, Anja, Jasna, Stanmanda, Pania, Jakki, Todd,** and **Paul**) for your various contributions to my scientific (and social) development. Thanks also for the favours done here and there. A special thanks goes to **Dr Jasna Rakonjac**. Thank you so much Jasna, for your encouragement, proofing/editing, and helpful discussions/ tutorials over the past couple of years. It has been greatly appreciated.

To anyone else who contributed to my thesis in some small way, e.g., lending equipment, handy tips, administration, "mental health morning teas", and stuff like that - thanks heaps. Thanks to my family and friends for all of your help (in several forms).

Finally to my husband **Pete**. Thank you for your support, patience, taking care of Gracie in the evenings, and frequently locking me in the office. You can go back to the shed and play now. XXXX.

RELATED PUBLICATIONS

Some of the material presented in this thesis has been published.

Forester, N.T., Parton, K., Lumsden, J.S., and O'Toole, P.W. (2000). Isolation of *Helicobacter mustelae* from ferrets in New Zealand. *New Zealand Veterinary Journal* **48**:65-69.

Forester, N., Lumsden, J.S., O'Croinin, T., and O'Toole, P.W. (2001). Sequence and antigenic variability of the *Helicobacter mustelae* surface ring protein Hsr. *Infection and Immunity* **69**(5):3447 – 3450.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 INTRODUCTION

1.1 The genus *Helicobacter*

1.1.1 General history

The first observations of spiral-shaped bacteria in animal stomachs were made by Rappin (1881), Bizzozero (1893) and Salomon (1896) (reported in Fox and Lee, 1997). Doenges was first to report a high incidence (103/242) of spiral bacteria in sections of gastric mucosa from human stomachs (Doenges, 1938). Steer and Colin-Jones (1975) described the presence of spiral organisms closely associated with the gastric mucosa from patients with gastritis while absent in non-inflamed stomach tissue. Warren and Marshall (1983) also observed these spiral bacteria associated with chronic inflamed gastritis, and noted a physical resemblance to *Campylobacter jejuni*. Subsequently, *Campylobacter* isolation techniques were employed to yield the first successful culture of the *Campylobacter*-like organisms (CLOs) from human antral biopsy specimens (Marshall and Warren, 1984). These organisms were classified as belonging to the *Campylobacter* genus until biochemical, physical, and finally phylogenetic 16S rRNA gene sequence analysis (Goodwin *et al.*, 1985; Jones *et al.*, 1985; and Romaniuk *et al.*, 1987) resulted in the creation of a new genus, *Helicobacter* (Goodwin *et al.*, 1989). To date more than 20 species of *Helicobacter* have been isolated from various hosts (Table 1.1), including *Helicobacter* species found in organs other than the stomach.

1.1.2 General characteristics of members of the genus *Helicobacter*

The Gram-negative *Helicobacters* exhibit varied morphological forms from spiral to rod shaped (3 x 0.5 μM). The cells are relatively slow growing (2 – 5 days), have relatively fastidious nutrient needs, and requires both a microaerobic atmosphere (5% CO_2, 90% N_2, 5% H_2), and a pH environment of pH 4.5 – 8.0 for successful culture. *Helicobacters* are motile, having one to several sheathed flagella per cell. A strong urease activity is a major property of *Helicobacters* associated with the host stomach, while most of those isolated from other organs do not generally exhibit urease activity (Table 1.1; Fox and Lee, 1997). Beyond the common characteristics, various

Helicobacter species presumably evolved specialised adaptive features in response to host natural selection.

Table 1.1 List of *Helicobacter* species and their hosts.

Species	Hosts	Primary site	Urease (+/-)	Other sites
H. pylori[a]	Human, macaque, cat	Stomach	+	
H. mustelae	Ferret, mink	Stomach	+	
H. felis[a]	Cat, dog	Stomach	+	
H. bizzozeronii[a,b]	Dog, human	Stomach	+	
"H. heilmannii"[a,b]	Dog, cat, human, monkey	Stomach	?	
H. nemestinae	Pig-tailed macaque	Stomach	+	
"H. suis"	Swine	Stomach	?	
H. acinonyx	Cheetah	Stomach	+	
"H. rappini"[a]	Sheep, dog, human, mice	Intestine	+	Liver (sheep), stomach
H. canis[a]	Dog, human	Intestine	-	Liver (dog)
H. hepaticus	Mice	Intestine	+	Liver
H. bilis	Mice, dog	Intestine	+	Liver, stomach (dog)
H. rodentium	Mice	Intestine	-	
H. trogontum	Rat	Intestine	+	
H. muridarum	Mice, rat	Intestine	+	Stomach (mice)
H. cinaedi[a]	Human, hamster	Intestine	-	
H. fennelliae	Human	Intestine	-	
H. pullorum[a]	Chicken, human	Intestine	-	Liver (chicken)
H. pametensis	Bird, swine	Intestine	-	
H. chlolecystus	Hamster	Liver	?	
H. suncus[c]	House musk shrew	Stomach	+	
H. winghamensis[d]	Human	Intestine	-	
H. aurati[e]	Syrian Hamster	Stomach	+	

Table data is mostly taken from Fox and Lee, 1997 (and references therein). [a]Some data suggest zoonotic potential; [b]Closely related, may be same species; [c]Goto *et al.*, 1998; [d]Melito *et al.*, 2001; [e]Patterson *et al.*, 2000.

1.1.3 Medical significance of *Helicobacters* to Humans

Helicobacter pylori cells inhabit the protective mucus layer covering the gastric epithelium, and a proportion are intimately associated with the gastric epithelium of susceptible human hosts (Blaser, 1993; Hessey *et al.*, 1990). Approximately 30% of the adult population in developed countries and 80 - 90% of the adult population in developing countries are infected by *Helicobacter pylori*, with the incidence of infection depending on factors such as age, socioeconomic status, and ethnicity (Dunn *et al*, 1997; Blaser, 1996). *H. pylori* is an important causative agent of peptic ulcer disease in humans (Marshall, 1995) and is also associated with other gastric diseases, ranging from mild, asymptomatic superficial gastritis to gastric carcinogenesis, all of which may take years to manifest (Lee *et al.*, 1993; Blaser, 1993, Forman, 1993). At the turn of the 20th century, stomach cancer was the leading cause of death in the USA, on average 32 deaths per 100,000 people. These rates have reduced over time as hygiene and living

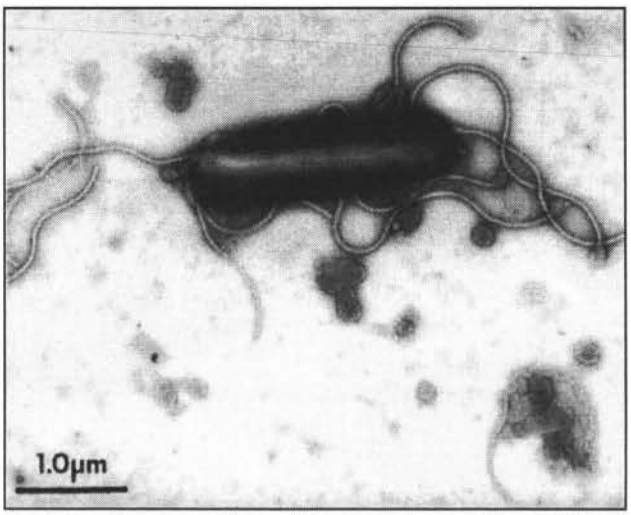
conditions have improved, although stomach cancers are still significantly represented in the cancer statistics (Blaser, 1996).

H. pylori induces a strong, yet relatively ineffectual immune response (Ferrero, 1997). These factors have motivated the intensive study of *Helicobacters*, in particular *H. pylori*, with many animal models employed to investigate various aspects of *Helicobacter* infection. These animal models include gnotobiotic piglets, primates, cats, dogs, ferrets and various rodents. Among these, the murine model has an advantage, since it can be infected with adapted *H. pylori* strains and there are many immunological reagents and mutant or transgenic strains available to exploit (Lee, 2000).

1.2 *Helicobacter mustelae*

1.2.1 General background

H. mustelae was the second member of the *Helicobacter* genus to be isolated. This species was cultured from the stomachs of ferrets (Section 1.2.2) (Fox *et al.*, 1986; Cave *et al.*, 1986; Fig. 1.1a). Up until 1997, *H. mustelae* had been isolated from stomachs of ferrets inhabiting America, England, Canada and Australia (Tompkins *et al.*, 1988; Fox and Lee, 1997). However, *H. mustelae* organisms were not detected in the stomachs of a group of New Zealand ferrets (Morris, *et al.*, 1988). Besides ferrets, *H. mustelae* has been isolated from the stomachs of other Mustelids (Section 1.2.2), including *Mustelae vison* (mink) (Fox *et al*, 1993), and most recently, *Mustela ermina* (stoat) (O'Toole and Forester, unpublished results), but has not been isolated from non-Mustelid animals. *H. mustelae* naturally infects ferrets, and virtually all ferrets with gastritis are infected with the bacterium (Fox and Lee, 1997). The disease induced by *H. mustelae* in ferrets mimics chronic diffuse (superficial) type gastritis in humans, caused by *Helicobacter pylori* (Fox *et al.*, 1990).



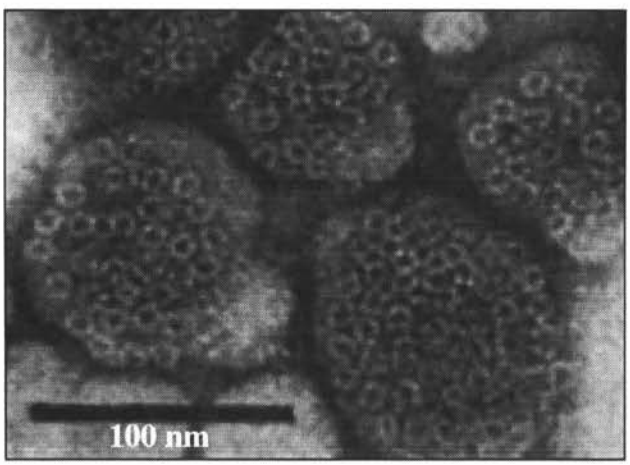


Figure 1.1 ***H. mustelae*** **cell morphology and Hsr protein rings**

Electron micrograph of *H. mustelae* cell morphology and flagellar configuration stained with 1% phosphotungstic acid. The figure is reproduced, with permission from the New Zealand Veterinary Association (copyright holder), from Forester *et al.*, 2000. B. High magnification electron micrograph of ammonium molybdate-stained membrane preparations showing the Hsr ring structures of *H. mustelae* strain 4298 (modified from O'Toole *et al.*, 1994).

1.2.2 Ferrets and the ferret animal model

Ferrets (*Mustela putorius furo*) belong to the subfamily of small carnivores (*Mustelinae*) within the *Mustelidae* family. Other members of the genus include weasels, stoats, mink, otters, badgers, skunks and wolverines (Fox, 1998). People have exploited ferrets for centuries, breeding for pelts, for hunting uses ("ferreting"), for biomedical research, and as cherished pets. In some countries ferrets are considered pests. For example, ferrets, along with stoats and weasels were released in New Zealand (1880s, Fox, 1998) in an attempt to reduce the population of rabbits, which were originally introduced as game, but bred uncontrollably in the absence of natural predators. However, the introduction of the new predators did not result in the decline of the rabbit population as expected, but have rather devastated the native bird population (Fox, 1998).

Ferrets are being used with an increasing frequency in almost all aspects of biomedical research, since the establishment of commercial breeding farms makes ferrets readily available. In addition, many similarities of the ferret anatomic, metabolic, and physiological features to those of humans, has seen the ferret being promoted as an animal model for disease in humans. The similarity of gastric anatomy, physiology and causative agent of gastric mucosal disease, lead to the suggested use of the ferret as an animal model for *Helicobacter*-induced pathogenesis and transmission (Cave *et al.*, 1986; Fox, *et al* 1992; O'Rourke *et al.*, 1992). Almost ten years later, this ferret model has largely been replaced by *Helicobacter*-adapted murine models (Section 1.1.3). Nevertheless, the ferret model has contributed valuable information, including the ability of *Helicobacters* to induce gastric adenocarcinoma (Fox *et al*, 1997), and allowed the testing of the effectiveness of antimicrobial chemotherapy (Otto *et al.*, 1990).

1.2.3 Characteristic features of *H. mustelae*

A general description of *Helicobacters* is given in Section 1.1.2. *H. mustelae* is morphologically distinct to *H. pylori* (Fox et al 1989). A summary of some important similarities and differences between *H. mustelae* and *H. pylori* are listed in Table 1.2.

Table 1.2 Characteristics of *H. mustelae* compared with *H. pylori*

Characteristic	H. mustelae	H. pylori
Culture/ Morphology		
Growth temperature, microaerobic conditions[a]	37°C, 42°C	37°C
Shape of Gram negative rods	slightly curved (2 x 0.5 μM)	spiral
Flagella (essential for colonisation)	sheathed	sheathed
Type and distribution of flagellae[b]	4-8 peritrichous	4-8 bipolar
Biochemistry		
Rapid urease activity[a]	+	+
Nitrate reduction[a]	+	-
Nalidixic acid (30μg disc)[a]	sensitive	resistant
Cephalothin (30μg disc)[a]	resistant	sensitive
Predominating fatty acids[a]	C16>C19>C14	C14>C19
Lipid A[f]	different structures	
DNA characteristics		
GC% (mol%)[b]	36	35-37
Proteins[c]		<40% similarity[c]
Urease (essential for colonisation)	+ (yes)	+ (yes)
Cytotoxin	-	+
8.5 nm surface rings (150 kDa monomer)	present	absent
Histology		
Cellular sites[d]	surface epithelium	intracellular junctions
Cell adherence "pedestals"[d] (%)	almost all	3-19%[d]
Dense inclusion bodies associated with the basal body of the flagellum[d]	+	-
Gastric Disease		
Naturally infected	yes	?
Inflammation – diffuse antral gastritis in host	+	+
Hypergastrinemia[e]	+	+
MALT lymphoma[g]	+	+

[a]Tompkins *et al.*, 1988. [b]Fox and Lee, 1997. [c]Morgan *et al.*, 1991. [d]O'Rourke *et al.*, 1992. [e]Perkins *et al.*, 1996. [f]Therisod *et al.*, 2001. [g]Erdman *et al.*, 1997.

O'Rourke and colleagues (1992) performed a detailed examination of the ultrastructure of *H. mustelae*. *H. mustelae* (Fig. 1.1a) is a short, slightly curved bacilli (2 x 0.5μM), with peritrichous sheathed flagella, and is found almost exclusively closely associated with the ferret gastric epithelium, forming adhesion "pedestals", with bacterium seldom found free in the overlying gastric mucus (in contrast to *H. pylori*). Significant numbers are found intracellularly, both endocytosed and invading. Coccoid forms have been detected in gastric sections with flagella grouped to one end. Dense inclusion bodies noted associated with the base of flagella were suggested to be involved in generation of energy for motility and cell wall formation (O'Rourke *et al.*, 1992). *H. mustelae* movement is distinct from other *Helicobacters*, with an efficient spinning motility rather than the classical movement of the spiral forms. This motility was thought to be linked to the aforementioned basal bodies and peritrichous flagellae, and may explain how *H. mustelae* achieved efficient movement through the thick epithelial mucus lining in the absence of spiral morphology (O'Rourke *et al.*, 1992). Another apparently unique

feature of *H. mustelae* is the *Helicobacter mustelae* surface ring protein (Hsr, Fig. 1.1; Section 1.3).

1.3 The *Helicobacter mustelae* surface ring protein (Hsr)

1.3.1 The Hsr protein

The Hsr protein was first observed, using electron microscopy, forming a regular array of 8-nm diameter ring-shaped structures covering the *H. mustelae* cell surface (O'Rouke *et al.*, 1992). The Hsr protein was further characterised using biochemical procedures by O'Toole *et al.* (1994). Protein monomers with an apparent molecular mass of 150 kDa were presumed to assemble together to form the 8.5 nm diameter ring-shaped structures. The number of monomers in a ring could not be accurately determined because of poor resolution at high magnification (Fig. 1.2). Each ring structure has an internal diameter of 4.25 nm and extends 6 nm from the outer membrane of the *Helicobacter mustelae* cells (O'Toole *et al.*, 1994). Loose association of many ring structures produced the regular arrays previously mentioned, and these Hsr rings constituted approximately 25% of the total cellular membrane proteins (O'Toole *et al.*, 1994). This collection of ring structures was originally suggested to be an S-layer (O'Rourke *et al.*, 1992), but was later shown to be distinct from classical S-layers (O'Toole *et al.*, 1994). The Hsr protein is most likely specific for *H. mustelae*, and has not been detected in *H. pylori* or *H. felis* (O'Toole *et al*, 1994).

Each Hsr monomer is produced as a preprotein, with a signal peptide of 47 amino acids, as determined by N-terminal amino acid sequencing (O'Toole *et al.*, 1994). At the time of discovery, it was considered unusually long for typical signal sequences from Gram-negative bacteria (O'Toole *et al.*, 1994; Pugsley 1993). The discovery of a group of proteins known as the autotransporters has suggested the mechanisms of signal peptide processing in these proteins (Section 1.3.2). The Hsr precursor is cleaved, presumably by a signal peptidase, to yield a mature Hsr protein (153 kDa) composed of 1472 amino acids, the general properties of which are listed in Table 1.3. Overall, the Hsr protein is large and generally hydrophilic.

Table 1.3 Summary of the amino acid characteristics of the Hsr protein

CHARACTERISTIC/ PROPERTY	
Preprotein	1519 aa
Molecular weight	158.0 kDa
Signal peptide[a]	47 aa
Molecular weight	5.4 kDa
Mature Hsr protein	1472 aa
Molecular weight	152.6 kDa
pI[a]	6.06
Charge at pH 7.0[a]	-4.19
Kyte & Doolittle average hydrophobicity score[a]	-3.06
Polar uncharged amino acids (C, G, N, Q, S, T, Y)[ab]	52.2%
Aliphatic sidechains (A, G, I, L, P, V)[b]	42.9%
Non polar/ Hydrophobic amino acids (A, F, I, L, M, P, V, W)[ab]	35.9%
Aromatic sidechains (F, W, Y)[ab]	8.5%
Basic (H, K, R)[ab]	6.1%
Acidic (D, E)[ab]	5.9%
Cys (1)[ac]	0.1%
Peripheral: Integral odds[a]	2.73

[a]O'Toole *et al*, 1994. [b]Letters in parentheses represent the single letter amino acid code. [c]Number in parentheses indicate the incidence of the item in the mature Hsr protein.

The Hsr protein is anchored in the outer membrane presumably via a hydrophobic β-barrel structure located in the C-terminal domain of the protein. This β-barrel modular structure is the common feature unifying autotransporter proteins (Henderson *et al.*, 1998) and is described in Section 1.3.2. Hsr lacks both the catalytic triad and appropriate cleavage sequence for proteolytic digestion, consistent with its retention on the cell surface (Forester *et al.*, 2001). The function of the Hsr protein is unknown, but it is speculated that the protein may be a potential virulence factor. It has been discounted as a haemagglutinin (Forester and O'Toole, unpublished previous work), and is unlikely to possess anti-phagocytic properties. The latter function was tested on a non-Hsr expressing variant of *H. mustelae* strain 4298, that spontaneously reverted to wild type expression upon passaging (O'Toole *et al*, 1994). The mechanism of this phase type variation is unclear. The protein has been shown to be a strong immunogen in rabbits (O'Toole *et al.*, 1994) and elicits antibody production in ferrets (Forester et al., 2000). The relative abundance of the Hsr protein in the outer membrane and exposure to the external environment make it an obvious target for host immune attack. Investigations of the function, efficient production, export and variability of this major surface antigen are ongoing.

1.3.2 Hsr belongs to the family of Autotransporter proteins

The Hsr protein has been proposed as a member of the expanding family (>40 members) of secreted autotransporter proteins from Gram-negative bacteria (Henderson *et al.*, 1998) (Table 1.4). Although striking sequence heterogeneity of all autotransporter proteins is evident, the converging feature of these proteins is the presence of the mechanism for efficient outer membrane localisation integrated into each of the secreted protein precursor, hence the descriptive term 'autotransporter'. The general structure of autotransporters comprises three functionally distinct domains: an amino-terminal signal peptide; the passenger (α) domain, conferring the major function of the protein; and the β domain, involved in pore formation and the mechanism of exporting the passenger domain to the external surface of the bacterium (Fig. 1.2). This mechanism was first described for the immunoglobulin A1 (IgA1) proteases (Klauser *et al.*, 1993).

Most autotransporter signal sequences resemble the Sec-dependent signal sequences, with a positively charged N domain, a hydrophobic H region and C-domain containing the signal peptidase cleavage site. These *sec*-dependent leader peptides are unusually long, with extended N-domains containing many charged amino acids, the suggested functions of which are to recruit accessory proteins involved in the secretion process, or to determine protein localisation (Henderson *et al.*, 1998). Autotransporter secretion employs the first step of the type II Sec-dependent secretion pathway (GSP, general secretory pathway) for export of the protein precursors from the cytoplasm to the periplasm (Pugsley, 1993, Hueck, 1998).

The carboxyl terminal β domain is assumed to form a pore by the spontaneous assembly of 10 – 18 (14 is most common) amphipathic 9-residue β-sheets in an antiparallel organisation, into a β-barrel conformation. Pore formation has been demonstrated for the *Bordetella pertussis* BrkA protein (Shannon and Fernandez, 1999). The extreme C-terminus of each protein shares a family signature motif, which is conserved and has suggested involvement in the efficient outer membrane localisation of the β-domain (Loveless and Saier, 1997). Interestingly, unlike other autotransporter proteins, the Hsr protein monomer has 34 amino acids following the end of the signature motif. The

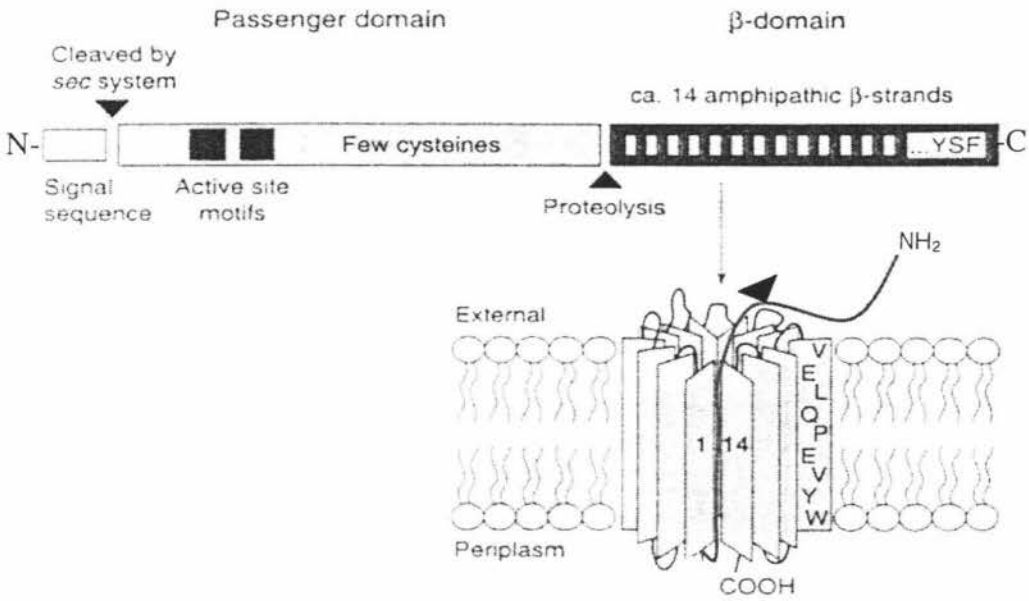


Figure 1.2 Autotransporter domains and membrane topology.

General structural organisation of the three structural domains and membrane topology of the β-domain of autotransporter proteins. **Top**: The three main domains are indicated (signal sequence, passenger domain, and β-domain). Each domain is described in the corresponding text. Proteolytic cleavage sites are indicated with arrowheads. Structural and functional motifs (e.g., active site motifs) are indicated. **Bottom**: The predicted membrane integration of the β-barrel structure is shown in relation to the outer membrane. This figure is a modified version of that which appears in Henderson *et al.*, 1998.

Table 1.4 Autotransporter proteins occur in many Gram-negative bacteria and have functions promoting disease in host organisms.

Protein functions	Microorganism (Autotransporter examples)
Adhesins	*Escherichia coli* (AIDA-I, Tsh, Ag43, TibA); *Bordetella pertussis* (Pertactins, TcfA); *Haemophilus influenzae* (Hia) *Helicobacter pylori* (AlpAB?)[a]
Proteases	*Neisseria gonorrhoeae*, *Neisseria meningitidis*, *H. influenzae* (IgA1 protease); *E. coli* (EspP, Pic); *Serratia marcescens* (Ssp, Ssp-h1, Ssp-h2); *Shigella flexneri* (SigA, Pic)
Invasins	*H. influenzae* (Hap)
Mediators of Motility	*S. flexneri* (IcsA)
Serum resistance	*B. pertussis* (BrkA)
Toxins	*E. coli* (EspC, Pet, Sat); *Helicobacter pylori* (VacA)
Unknown	*Helicobacter mustelae* (Hsr); *H. influenzae* (Hsf); *S. flexneri* (SepA, ShMu); *Rickettsia ssp* (rOmpA, rOmpB; SIpT)

[a]Odenbreit *et al.*, 1999

significance of this has not been investigated.

The passenger (α) domain contains the least conserved amino acid sequence of the autotransporter protein domains. Sequence divergence is attributed to the range of diverse functions of each protein, which include: adhesins, proteases, toxins, invasins, motility mediators, and serum resistance (Henderson *et al.*, 1998). The transfer of the passenger domain transfer to the outer surface of the bacterium is often followed by specified proteolytic release of this domain into the external milieu. The proteolytic determinants, if present, are frequently carried in the passenger domain itself, thus leading to autoproteolysis. The exceptions are the adhesins; AIDA-I protein, *E. coli* Ag43, and the *B. pertussis* pertactin proteins, which may require a separate membrane-associated protease for cleavage (Henderson *et al.*, 1998). Interestingly, in this group of proteins, the passenger domain remains associated with the membrane bound β-domain after proteolysis of the peptide chain between them. The Hsr protein does not undergo proteolysis and its equivalent α-domain is exposed to the surface and presumably anchored by the β-domain (O'Toole *et al.*, 1994). A noteworthy feature of the passenger domain is the conserved scarcity of cysteine residues. Formation of disulphide bonds in the passenger domain have been shown to prevent or reduce (three-fold) its localisation to the outer membrane (Jose *et al.*, 1996; Veiga *et al.*, 1999). These results imply that extensive tertiary structure impairs efficient translocation of the α-domain through the proposed β-barrel pore. The minimal complete translocation unit (TU) has been demonstrated to include the β-domain and a linker region extending into the α-domain of the AIDA-I protein (Maurer *et al.*, 1999). It appears likely that the complete TU functions have been conserved across autotransporters.

The current model of autotransporter secretion is as follows: The nascent polypeptide chain is exported from the cytoplasm to the periplasm in a Sec-dependent manner. The C-terminal residues integrate into the outer membrane, followed by spontaneous pore formation via membrane spanning antiparallel amphipathic β-sheets, and finally translocation of the passenger domain onto the outer surface of the bacterium. Threading of the 'unfolded' N-terminal passenger domain through the β-barrel pore is thought to occur via a hairpin-like structure, which is most likely guided by the linking region.

The autotransporter proteins have been phylogenetically separated by DNA and protein homology into individual subfamilies (Henderson *et al.*, 1998). The Hsr passenger domain has no significant homology to other autotransporters. Hsr lacks the consensus serine protease active site motif (GDSGSP), consistent with its retention on the cell surface. It does not have RGD motifs associated with attachment to the plasma membrane of mammalian cells (D'Souza, *et al.*, 1991). Consequently, the Hsr protein has been singled out as the single protein in the Hsr family or family III (Henderson *et al.*, 1998).

1.3.3 The *hsr* gene and generation of the sequence of the 12 kb *hsr* locus – preliminary information

The sequence of the 4557 bp *hsr* and upstream elements was previously identified, isolated, and sequenced by O'Toole *et al.* (1994). The area housing the upstream elements was scrutinised for –10 and –35 promoter sequences, and was apparently expressed from non-typical upstream transcription/translation elements. Re-examination of these *hsr* upstream elements, as part of a study into the efficient Hsr export, lead to the discovery of a DNA rearrangement introducing 22 bp of EMBL3 cloning vector DNA and other unknown sequence into the immediate upstream of the *hsr* gene sequence (Fig. 1.3). This insertion effectively removed the native ribosome binding site and transcription start site, and was assumed to have occurred during the generation of the EMBL3 lambda clone, λE2 (Fig. 1.3). Consequently, a new *Hm4298* genomic library was prepared and screened for Hsr expression, using polyclonal antibody to Hsr (JA6, O'Toole *et al,* 1994). The lambda clone λP1C was expanded and mapped by restriction enzyme analysis. Strategically placed λP1C sub-fragments were then subcloned into general cloning vectors for DNA sequence analysis (Figure 1.3). DNA fragments containing the native *hsr* upstream DNA sequence were apparently lethal using a number of different cloning strategies. However, the sequence of neighbouring clones allowed for the identification of primers that could be used to amplify connecting DNA fragment from λP1C DNA that contained the putative upstream elements. Confirmation of this DNA sequence was performed using *H. mustelae* strain *4298* genomic DNA as the PCR template in place of the λP1C derived sequences (Forester and O'Toole, previous unpublished work).

DNA sequencing of this fragment resulted in the identification of a consensus ribosome binding site (AGGAGA) seven nucleotides upstream of the translational start site. Primer extension analysis was performed on strain *4298* total RNA. The transcription initiation site was located 56 bp upstream of the translation start site (O'Toole and Forester, unpublished).

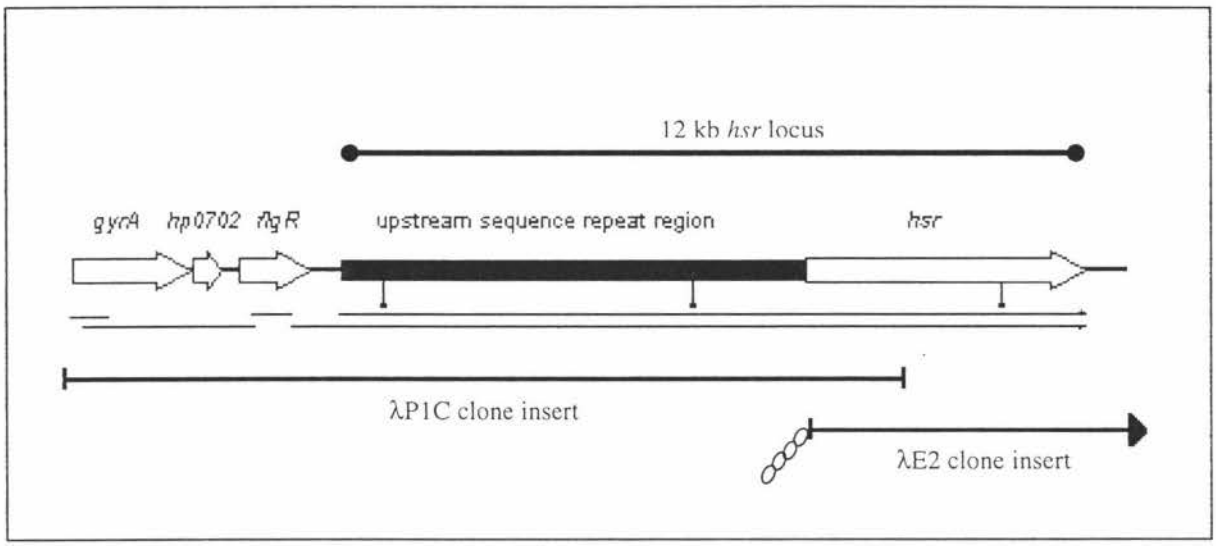


Figure 1.3 Genomic organisation of the *hsr* locus of *H. mustelae* strain 4298.

The genetic arrangement of the *hsr* locus (knobbed line) is shown. Putative genes (unfilled arrows) neighbouring the repeat region (black box) are labelled according to the most related protein according to BLAST protein and nucleotide homology searches. The gap between the descending marker lines denotes 5000 bp of DNA sequence. The horizontal lines directly below the markers represent a summary of the DNA sequencing performed in this area of the strain 4298 genome, with the top line + strand and bottom - strand with respect to the *hsr* gene. The two indicated lambda clones, λP1C and λE2, are indicated below representative lines depicting the regions carried within each clone insert. The λE2 clone is 'arrowed' to show a continuation of sequence beyond the end of the *hsr* gene. The looped region shown at the 5' end of the λE2 clone represents DNA that was scrambled during the generation of the original EMBL3 lambda library prior to the original sequencing project.

DNA sequencing of other subclones of λP1C insert surprisingly revealing DNA sequence containing previously unidentified *hsr*-like repeat sequences upstream of the *hsr* gene. Sequencing 5' of the *hsr* gene, to the nearest gene neighbour, resulted in a region of approximately 7.0 kb of upstream *hsr*-like sequence being detected (Forester,

previous work). A total of approximately 12 kb (Fig. 1.3) constituted an area named the *hsr* locus (HSRL), which included the *hsr* gene plus the upstream repeats, and provided the starting material for this study.

1.4 Functional roles of surface exposed proteins of Gram-negative mucosal pathogens

Surface exposed bacteria are in constant contact with the external host environment. The Majority of these surface proteins contribute, at least in part, to bacterial pathogenicity. Abundantly expressed S-layer proteins (SLPs) of Gram-negative bacteria are involved in bacterial protection and virulence. For example, *Campylobacter fetus*, *Aeromonas salmonicida*, *A. serpens, and Caulobacter crescentus* SLPs shield attack from bacterial parasites, such as *Bdellovibrio bacteriovorus* (reviewed in Sara and Sleytr, 2000). The *C. fetus* Sap proteins and *A. salmonicida* A-layer prevent complement binding (serum resistance) and both proteins exhibit antigenic variation (Section 1.5.2). In addition, post-translational modification (tyrosine phosphorylation) of the AhsA S-layer protein of the *Aeromonas hydrophila* has been implicated in sensing and signalling events (Thomas and Trust, 1995). Other functions have been described for SLPs in other bacteria from distinct ecological niches, but not an integrative function for all (Boot and Pouwels, 1996; Sara and Sleytr, 2000). Surface exposed autotransporter proteins, like SLPs, are implicated in bacterial virulence via a broad spectrum of adaptive functions (refer to Section 1.3.2).

Analysis of the genome of *H. pylori*, the closest sequenced relative to *H. mustelae*, identified several outer membrane proteins demonstrating a large range of functions for cell maintenance within the host stomach (Tomb *et al.*, 1997; Alm *et al.*, 1999). These functions include porins / adhesins (colonisation), siderophores (iron uptake), antigenic variation (colonisation / persistence), transporters, pilin (colonisation), flagella components (motility), toxins, restriction modification components (protection), and proteases (evasion). Often, more than one functional activity is presented in a single protein. For example, the *H. pylori* VacA autotransporter has roles in gastric epithelial cell toxicity and antigenic variation (Atherton, *et al.*, 1995; Cover and Blaser, 1992). The more distantly related *Campylobacter jejuni* has distinct surface properties from

those of *H. pylori,* with large numbers of surface polysaccharides and lipooligosaccharide structures predominating over surface proteins (Parkhill *et al,* 2000a), but serving similar functions associated with evasion of host immune surveillance. Despite diverse surface structures, the overall functions of surface proteins are conserved in most Gram-negative mucosal pathogens.

1.5 Generation of bacterial protein variability

1.5.1 Repeats sequences in prokaryotes

Repeat sequences are ubiquitously present in the genomes of all organisms. Genomic sequencing has facilitated the detection of repeat sequences. Clues from the genomic incidence and positioning of sequence repeats are unravelling the functions of these repeat sequences together with the molecular mechanisms that generate them. Seven distinct repeat types were detected in the meningococcal genome, comprising: insertion sequence (IS) elements, repetitive extragenic palindromic sequences (REPs), AT rich repeats (ATR), Correia elements (CE), neisserial intergenic mosaic elements (NIMEs), prophage sequences, and DNA uptake sequence (Parkhill *et al.*, 2000b). The 10-bp DNA uptake recognition sequence (5'-GCCGTCTGAA-3') is abundant in the *Neiserria* genome (almost 2000 times) functioning in recognition of homospecific DNA during natural transformation. Prokaryotic interspersed repetitive DNA has important roles in chromosome structure, regulating gene expression and can contribute to virulence (Lupski and Weinstock, 1992). A single type of repetitive sequences, the stem loop forming REP sequences of *S. typhimurium* and *E. coli*, have been implicated in gene regulation (mRNA stabiliser, differential translation (Newbury *et al.*, 1987)), chromosome folding and chromosome rearrangements (Stern *et al*, 1984).

Short tandemly repeated DNA sequences are divided into three general classes. Homogeneous repeats comprise homopolymeric runs of a single nucleotide or short sequence repeats 1-6 nucleotides in length. Heterogeneous repeats consist of mixed homogeneous repeats, while degenerate repeats are tandem repeats with small nucleotide changes in the consensus repeat sequence (van Belkum *et al.*, 1998). Short sequence repeats have roles in bacterial phase variation (Section 1.5.2) and gene

regulation mediated by the formation of hairpin structures. Degeneracy in tandem repeats provides a mechanism for antigenic variability in *H. influenzae,* the longer the repeat, the higher the incidence of sequence heterogeneity (van Belkum, *et al.*, 1997). Large repeat sequence arrays (1 - >2001 kb) are almost exclusively involved in cell-surface functions in *N. meningitidis.* Repeat arrays were found flanking the genes of outer membrane porins (e.g. *porAB, lbpAB, tbpAB* and *hpuAB*) (Parkhill *et al.*, 2000b). Repeat sequences thus far have been associated with genes involved in bacterial pathogenesis and when present around genes can be positive indicators of potential virulence factors.

1.5.2 Pathoadaptive DNA rearrangements and mutations

The adaptive host immune system functions to eliminate or restrict pathogen replication. In the process it selects for pathogens with the ability to evade host defences and replicate, thus persisting within a population of susceptible or at least partially susceptible hosts (Brunham *et al.*, 1993). Small apparently non-programmed DNA changes, such as point mutations, small insertions and/or deletions in DNA sequences facilitate antigenic drift of a protein through natural selection and can lead to increased bacterial infectivity. These mutations may also confer a selective advantage, in appropriate conditions, by changing (modification or loss) a function of protein enhancing pathogenesis without gaining DNA encoding for specific virulence factors (Sokurenko *et al.*, 1999).

Examinations of genome sequences have noted that antigenic variation is a common feature among mucosal pathogens (Wren 2000). Antigenic variation comprises DNA alterations, generally associated with host immune evasion and, less frequently, niche-determined adaptative features such as tissue tropism (Seifert and So, 1988; Conner *et al.*, 1998). Strictly speaking, antigenic variation arises from a single bacterial strain that is capable of expressing several variants of a cellular component, not due to unprogrammed background mutation (Seifert and So, 1988). However, homologous recombinations between cellular DNA and donor DNA taken up during autolysis of neighbouring cells have been included in the description of antigenic variation related mechanisms (antigenic shift) of some Gram-negative bacteria (Gilsdorf, 1998;).

Horizontal DNA transfer, leading to a gain in antigenic diversity/ function, occurs upon transposition into an expressed gene sequences (Seifert *et al*, 1988). Phase variation mostly involves a simple "on" or "off" switching of expression of the particular protein, and is viewed as distinct from, but a subset of, the mechanisms related to antigenic variation (reviewed in Henderson, *et al.*, 1999).

Phase variation is widespread in Gram-negative bacteria, with roles in immune evasion of surface components, adhesin shedding, sensor / regulator systems and restriction/ modification. The function of phase variation is thought to be to provide a 'rapid' and reversible strategy of phenotype switching that permits the survival of the bacterial population in more than one environment. Phase variation mechanisms include site specific DNA inversions, recombination, slipped strand mispairing of small repetitive DNA during replication, and epigenetic variation involving differential methylation of DNA (Henderson *et al.*, 1999; Borst and Greaves, 1987). Site-specific DNA inversion is described for the type I fimbriae of *Escherichia coli* (Blomfield *et al.*, 1997), the flagellar of *Salmonella typhimurium* (Komano, 1999), and SlpAB proteins of *Lactobacillus acidopilus* (Boot, *et al.*, 1996b). Phase variation-related recombination is highlighted by the generation of non-piliated L-pilin and S-pilin variants of *N. gonorrhoeae* (Haas and Meyer, 1986; Haas *et al.*, 1987; Gibbs *et al.*, 1989). Slipped strand mispairing (SSM) during replication of small tandem repeats can lead to two functional outcomes depending on the placement of the slipped element. Truncated or aberrant frameshift products can be produced when the SSM occurs within in a gene, particularly for dimeric, tetrameric and pentameric repeats. Examples of this type of mechanism are the Lipopolysaccharide (LPS) of *H. pylori* (Marshall *et al.*, 1998), LPS biosynthesis proteins of *H. influenzae* (Weiser *et al.*, 1990), and the opacity protein (Opa) of *N. gonorrhoeae* (Meyer *et al.*, 1990). Strand slippage of small sequence repeats (SSRs) in proximity to promoter or other regulatory elements usually results in a more complexed expression of the cellular component than the simple OFF and ON switch. This type of volume-controlled phase variation mechanism occurs in the LKP-fimbriae of *H. influenzae*, the PorA porin (van der Ende, *et al.*, 1995) and the Opc adhesin of *N. meningitidis* (Henderson *et al.*, 1999). The amount of cell component expression is dependent on the resulting spatial arrangement of upstream elements after SSM has occurred. Of the sequenced genomes, *N. meningitidis* encodes 142 potentially phase-variable proteins, more than any other pathogen to date (Tettelin *et al.*, 2000).

Phase variation may also be achieved by modification of the original genotype. Differential methylation of GATC sequence is involved in the expression of antigen 43 autotransporter protein of *E. coli* ((Ag43) Henderson *et al.*, 1997; Owen *et al.*, 1996). All systems employ a number of auxiliary proteins, and switching of phenotypes appears to be globally regulated (Henderson *et al.*, 1999).

Other recombination-based mechanisms of antigenic variation include homologous recombination, DNA inversion, gene amplification, gene conversion, and horizontal DNA transfer. The mechanisms employed differ according to individual gene arrangements all resulting in antigenic variability. *N. gonorrhoeae* pilin antigenic variation occurs by multiple *recA*-dependent unidirectional gene conversions of partial silent genes in pilS loci to the functionally expressed recipient *pilE* loci via a pilE/pilS hybrid molecule, most likely following replication of recipient *pilE* (Zhang *et al.*, Mehr and Seifert, 1998; Hamrick *et al.*, 2001; Howell-Adams and Seifert, 1999; Howell-Adams and Seifert, 2000). This is the only description of a molecular mechanism of gene conversion for a bacterium to date (Howell-Adams and Seifert, 2000). Similar multiple gene conversion events are presumed to occur for the *Mycoplasma synoviae vlhA* haemagglutinin gene (Noormohammadi *et al.*, 2000).

Antigenic variation and phase variation of the surface layer proteins (*Sap*) of *C. fetus* is mediated by DNA inversion recombination (not site-specific) of a promoter sequence in located within a 6.2-kb invertible element (Dworkin and Blaser, 1996). *recA*-dependent DNA inversion recombination 'delivers' the promoter sequence to one of multiple silent genes, independent of the distance between inversion sites (Dworkin *et al.*, 1997; Dworkin and Blaser, 1997a; reviewed in Dworkin and Blaser 1997b). Likewise, a site-specific DNA-inversion model has been described for the control of Omp1 major surface protein phase and antigenic variation in *Dichelobacter nodosus* (Moses, *et al.*, 1995). Silent copies of variable major protein of *Borrelia hermsii* are located on storage plasmids. Site specific recombination between silent plasmid and expression plasmid *vmp* gene result novel structural gene in the plasmid expression site (Plasterk, *et al.*, 1985).

Horizontal DNA transfer can be carried out by transformation, conjugation or transduction. Horizontal transfer is a common feature of naturally transformable

mucosae-inhabiting Gram negative bacteria such as *Haemophilus influenzae, Helicobacter pylori, Neisseria gonorrhoeae,* and *H. mustelae. H. influenzae*, *N. gonorrhoeae*, and *N. menigitidis*. Transformation, in these bacteria, is dependent on invariant DNA uptake sequences (9 or 10 bp) (Gilsdorf, 1998; Parkhill *et al.*, 2000b). These sequences appear randomly distributed across the respective genomes. However, these sequences have not been identified for *Helicobacter pylori* (Section 1.5.3). Horizontal transfer between mixed strains of *Helicobacter* has been demonstrated during co-colonisation (Kersulyte *et al.*, 1999), although it is not known whether transformation or a conjugation-like mechanism is employed (Marshall *et al.*, 1998). In addition, horizontal DNA transfer (Lateral transfer) between different bacteria has been demonstrated to occur. For example, Pathogenicity islands (PAIs) carry assemblies of genes implicated in virulence functions and are associated with more virulent variants of a species. PAIs can spread through bacterial populations by horizontal gene transfer (e.g. *E. coli, Yersinnia* spp, *Heliobacter pylori*) (Hacker *et al.*, 1997).

Frequent gene amplifications are observed of the *H. influenzae* type b capsule (Cap b locus) Hib capsular genes in chromosome region characterised by two tandem 18-kb repeats and up to a max of 5 identified (Corn *et al.*, 1993). The quantity of capsule is proportional to number of tandem repeats. Loss of capsule expression is attributed to *rec*-dependent recombination between repeats deleting a region containing essential genes for capsule expression (Hoseith, *et al.*, 1986).

1.5.3 *Helicobacter* natural competence

Very little is known about how DNA is taken up during transformation of the naturally competent *Helicobacter* species. Naturally transformable *H. pylori* lacks a transformation-targeting system like those found associated with *Neisseria* DNA uptake recognition sequences (Saunders *et al.*, 1999). Horizontal DNA transfer occurs frequently between *H. pylori* strains during mixed infections (Kersulyte *et al.*, 1999). Characterised components essential for natural transformation of *H. pylori* to date comprise proteins encoded by the *comB* operon (Hofreuter *et al.*, 1998), *rdpA* (Smeets *et al.*, 2000), and *comH* (Smeets *et al.*, 2001. Homologues of the *comH* and *comB* were not detected in *H. mustelae* or *H. felis* genomes by hybridisation studies (Hofreuter *et*

al., 1998; Smeets *et al.*, 2001). These results suggested a lack of conservation of transformation genes, and hence mechanisms of DNA transformation in other naturally competent *Helicobacter* species (Smeets *et al.*, 2001).

1.6 Preliminary research objectives

Investigations into the efficient export of the Hsr protein to the external surface of *Helicobacter mustelae* cells have revealed extensive *hsr*-like repeats located in the upstream DNA sequence of the *hsr* gene and has defined an *hsr* locus of 12 kb. These findings have added another dimension to the Hsr story and warrant investigation to gather evidence for the function of the *H. mustelae* surface ring protein in *H. mustelae* infection in ferrets. The surface exposure of this abundantly expressed protein makes it a distinct target for host immune selection. A notion that the upstream repeats could function as a reservoir for potential sequence variation in the *hsr* gene sequence was used as a starting concept to direct experiments.

The principal objective of this study was to analyse the uncharacterised features of the 12 kb *hsr* locus of the *H. mustelae* strain 4298, in particular, to find out the extent and purpose of the recently discovered *hsr*-related repeat DNA sequences. Use of computer analysis software would be initially utilised to identify and organise all DNA repeats in the *hsr* locus. The information gathered would be used to identify key regions of the *hsr* gene homologous to the repeat sequences. These regions, thus identified, may yield clues as to the purpose of the *hsr* locus repeat sequences and the Hsr protein domain comprising them. Comparisons of the corresponding regions in different strains may further facilitate the determination of repeat function. A secondary aim was to identify putative open reading frames within the *hsr* locus and their relationship to the *hsr* gene and to Hsr production, using knockout mutagenesis and subsequent protein analysis tools.

During the course of the study, objectives were appropriately adjusted to accommodate expanding goals that were determined by the findings of the preliminary experiments.

2.0 MATERIALS AND METHODS

2.1 Bacterial strains, culture, and storage conditions

2.1.1 Bacterial strains

Bacterial strains used in this study are listed in Table 2.1.

Table 2.1. Bacterial strains used in this study.

Bacterial strain	Characteristics or genotype	Source / reference
E. coli ER2206	*endA1 thiI supE44 mcr67(mcrA-) Δ(mcrBC-hsdRMS-mrr)114::IS10 (lac) U169/F' proAB lacIqZΔM15 Tn10* (Tet^r)	New England Biolabs
H. mustelae 4298	Laboratory-passaged strain (ferret isolate)	J.G. Fox, Division of Comparative Medicine, Massacusetts Institute of Technology, Cambridge, Ma., USA
H .mustelae F6	New Zealand colony raised ferret isolate	This study
H. mustelae F7	New Zealand colony raised ferret isolate	This study
H. mustelae F8	New Zealand wild ferret isolate	This study
H. mustelae F11	New Zealand wild ferret isolate	This study
H. mustelae F15	New Zealand wild ferret isolate	This study
H. mustelae F21	New Zealand colony raised ferret isolate	This study

Escherichia coli was grown at 37°C on Luria-Bertani agar (LBA, Section 2.2), or in Luria-Betani broth (LB, Section 2.2) with shaking at approximately 200 rpm. Frozen stocks were kept at –70°C in Luria-Bertani broth (LB, Section 2.2) containing 20% glycerol.

Helicobacter cultures were grown on Chocolate blood agar (CBA, Section 2.2), or Columbia serum agar (CSA, Section 2.2) in an atmosphere containing 5% CO_2, provided by a CO_2 incubator (Revco Scientific) at 37°C. Alternatively, liquid cultures were grown in Brain Heart Infusion broth (BHI, Section 2.2) supplemented with 5% sterile heat-treated horse serum in microaerobic conditions generated by CampyGen sachets (Oxoid). Horse serum was obtained commercially (Life Technologies) and

complement inactivated prior to use by incubation at 60°C for at least 30 min after equilibrating to temperature. Culture stocks of cells were made by aseptically harvesting a plate of a two-day-old culture into sterile BHI/Glycerol mix [1 x BHI/20% glycerol] and kept at -70°C for long term storage.

2.1.2 *Helicobacter* isolation from ferret stomachs

2.1.2.1 Processing of ferret stomachs

The stomachs of 21 New Zealand ferrets were obtained from two separate sources. Four ferrets were from a breeding colony being sacrificed as part of an investigation into the cause of an illness unrelated to the ferret stomach. The remaining 17 ferrets were gathered from the lower Hawkes Bay region and trapped as part of an Agriquality New Zealand study, to assess ferrets as reliable indicator animals for the early diagnosis of livestock tuberculosis. The stomachs were dissected from the ferret carcasses prior to delivery. The time from stomach collection until processing ranged from an hour to at least three days depending on how long before traps were checked. The body condition and sex of each ferret was assessed when stomachs were collected.

Ferret stomachs were rinsed with sterile distilled water before opening with a clean sterile scalpel to release and discard the contents of the stomach. The exposed tissue was rinsed with sterile distilled water and the excess mucus was removed with a sterile loop. An area from the middle of the stomach (spanning all divisions of the stomach) was scraped firmly with a sterile loop to collect cells beneath the mucus layer and transferred into an eppendorf tube containing 1ml of BHI broth. Half of the sample of the tissue was pelleted at 20,800 x g, 1 min at room temperature (RT) and the pellet was stored at –20 °C until required for DNA extraction (Section 2.6.3). The remainder of the stomach was kept in buffered formalin [10% formalin (v/v), 3.98 g/L $NaH_2PO_4.H_20$, 6.4 g/L Na_2HPO_4] at RT.

Serial 10-fold dilutions of the tissue suspension were prepared and 50 µl of each was spread in duplicate onto CBA plates up to 10^{-6} dilution. The sample plates were incubated at 37°C in 5% CO_2 for 5 - 7 days until colonies began to appear. Colonies

that matched the morphological appearance of *Helicobacter* were streaked onto fresh CBA plates and incubated as above for 2 - 3 days. The 2 - 3 day old cells were tested for urease, oxidase and catalase activity (Section 2.1.2.2). Those that were positive in all three tests were checked by phase contrast microscopy (Section 2.12) for cellular morphology and characteristic *Helicobacter* motility. One representative isolate was chosen for further investigations.

2.1.2.2 Biochemical analysis of putative Helicobacter isolates cultured from ferret stomachs

For determination of urease activity, part of a test colony was suspended in 50 µl of filter-sterilised urease medium [1% peptone, 5% NaCl, 1% glucose, 2% KH_2P0_4, 0.012% Phenol red, 2% (w/v) Urea, pH6.8]. Urease-positive cultures changed the urease medium from an orange colour to a vivid pink/red colour.

For assessment of oxidase activity, part of a test colony (taking care not to carry over any blood agar) was smeared onto a piece of filter paper slightly dampened with freshly made oxidase reagent [~1% aqueous tetramethyl p-phenylenediamine dihydrochlorate]. A deep purple colour (within ~15 seconds) was produced by oxidase positive colonies.

2.2 Media and additives

Media used in this study are listed in Table 2.2. All media were made up to volume using water purified by the Milli-Q Reagent Water System (Millipore). Sterilisation of media was either by autoclaving at 121°C for 20 minutes with 15 pounds pressure or by passing media, or media components through a 0.2 µM or a 0.4 µM porosity sterile filter in to a pre-sterilised vessel. Supplements and antibiotics added aseptically to the growth media (≤50°C) when required.

Table 2.2. Recipes for media used in this study.

Medium	Recipe	Reference / supplier
Brain heart infusion broth (BHI)	Brain heart infusion broth (Oxoid) made to 95% volume with water as per supplier directions. Sterilise.	Oxoid
Columbia agar base (CA)	Columbia agar base made to 95% volume with water as per supplier directions. Sterilise. Store RT and microwave heat for later use.	Difco or Oxoid
Columbia blood agar (CBA)	Sterile CA cooled to 70°C, add 5% (final concentration) defibrinated horse blood (Life Technologies), cook at 70°C for 20 minutes for haemolysis (chocolate brown appearance) to occur and complement inactivation. Cool to approximately 50°C and pour.	Difco or Oxoid
Columbia serum agar (CSA)	Sterile CA cooled to 50°C, add 5% (final concentration) defibrinated horse serum (complement inactivated 30 minutes at 60°C) and pour.	Difco or Oxoid
Luria-Bertani agar (LBA)	LB with 1.5% (w/v) bacto-agar (Oxoid). Sterilise, cool to approximately 50°C and pour.	Sambrook *et al.*, 1989
Luria-Bertani broth (LB)	1% (w/v) tryptone (Difco), 0.5% (w/v) yeast extract (Merck), 0.5% NaCl; pH 7.0 with NaOH.	Sambrook *et al.*, 1989

Kanamycin sulphate was used for selection of clones possessing resistance gene, *aph3A*. Titration of kanamycin concentration from two commercial sources (Roche and Sigma) was required during the course of this study, due to poor selectivity at initial kanamycin concentrations of 20 µg/ml usually used for selection of *E. coli* clones. Dilutions of the sterile kanamycin stock solution were performed to the following final concentrations in CBA agar plates: 10; 20; 25; 50; 80; 100; 160 µg/ml. 100 µl of a suspension of *H. mustelae* strain 4298 and *E. coli* ER2206 at approximately OD(600) ~1.5 was spread and grown for 2 days under appropriate growth conditions (Section 2.1). For Roche kanamycin sulphate, the cells died between 100 and 160 µ/ml. For Sigma, the cells died between 50 and 80 µg/ml. Therefore, 60 µg/ml final concentration of Sigma antibiotic was used in all the respective experiments.

2.3 Oligonucleotide primers

Oligonucleotides used during this study are listed in Table 2.3. Lyophilised primers were obtained commercially and reconstituted in 200 µl sterile PCR grade Milli-Q water.

Table 2.3. Oligonucleotide primers used in this study.

Primer	Sequence 5'→3'	Priming region	Application in this study / Section reference
HS16sF	aggctatgacgggtatccggc	Helicobacter-specific 16S rRNA sequence	16S rRNA gene sequence determination, Section 3.2.1
HS16sR	ggtctagcaagctagacactc c	Helicobacter-specific 16S rRNA sequence	16S rRNA gene sequence determination, Section 3.2.1
Kan2	gcgaaccatttgaggtgatag g	Start of aphA3 gene +	PCR check for aphA3 insertion into orf2 of strain 4298, Section 3.5.3
ntf005	attgctactactcttgcattttc	HSRL+ 7575 – 97 HSRL- 1736 – 58; 3376 - 98; 13154 – 76	*hsr* variability (repeat region), Section 3.2.4.1
ntf008	tttcgatcgagtctcaataagg cgtatgg	HSRL+ 11016 – 33	*hsr* variability (β-domain), Section 3.2.4.1
ntf009	tttcgatcgatcgatgtttgcta ggtag	HSRL- 10997 – 1015	*hsr* variability (central region), Section 3.2.4.1
ntf011	tttagatctagtctcaataaggc gtatgg	HSRL+ 11016 – 33	PCR check of the 3' sequence of the *hsr* gene, Section 3.3.5
ntf016	gccccagagtggctttgtgc	HSRL + 6058 – 77	Check for aphA3 insertion in orf2 of strain 4298, Section 3.5.3
ntf026	gaacttctagccctggcagc	HSRL+ 8977 – 96; 9823 – 42 HSRL- 6 – 26; 719 – 39; 4247 – 67; 2625 - 44	*hsr* variability (central region), Section 3.2.4.1
ntf031	tttggatcctcaaatgccgcgc cgtg	HSRL + 7330 – 48	Probe amplification, Section 3.2.3
ntf034	cattgattacaggccctccgc	HSRL+ 958 – 77; 6533 – 52; 13805 – 24; HSRL- 8065 - 84	pUC19-λE2/E1 DNA sequencing, Section 3.3.4; Probe amplification, Section 3.2.3; PCR check of the 3' sequence of the *hsr* gene, Section 3.3.5; *hsr* variability (repeat region), Section 3.2.4.1
ntf037	cccctcagatctccctcgtgc ttacttcagc	HSRL- 12082 –101	*hsr* variability (β-domain), Section 3.2.4.1; PCR check of the 3' sequence of the *hsr* gene, Section 3.3.5
ntf050	cgtttccagtaagcaaaatgg	HSRL+ 14790 – 810 HSRL- 8899 - 919	pUC19-λE2/E2 sequencing, Section 3.3.5
ntf051	gcacctgtttgagtctagg	HSRL+ 12318-35	pUC19-λE2/E1 sequencing, Section 3.3.4
ntf052	gccatgcacaaagccgcg	3' HSRL sequence -	pUC19-λE2/E2sequencing, Section 3.3.5; PCR check of the 3' sequence of the *hsr* gene, Section 3.3.5
ntf053	tgctagtagtggccctgc	HSRL+ 13939 -56	pUC19-λE2/E1 sequencing, Section 3.3.4
ntf054	gacatgcaataaagcagcgc	3' HSRL sequence -	pUC19-λE2/E2sequencing, Section 3.3.5
ntf055	cgctcttggcgcaaaattgc	HSRL- 12728 – 47	pUC19-λE2/E1 sequencing, Section 3.3.4; PCR check of the 3' sequence of the *hsr* gene, Section 3.3.5
ntf056	gcatcccggggagctatcac	HSRL- 18488 – 67; 3488 – 507; 13265 – 84	pUC19-λE2/E1 sequencing, Section 3.3.4
ntf057	gtgatagctccccgggatgc	HSRL+ 18488 – 67; 3488 –507; 13265 – 84	pUC19-λE2/E1 sequencing, Section 3.3.4 PCR check of the 3' sequence of the *hsr* gene, Section 3.3.5
ntf064	ccgcagcaccagtaacagc	HSRL- 14614 - 32	Sequencing primer; Section 3.3.5
pUC/M13 FP	gttttcccagtcacgac	PUC18/19 +/-	pUC19-λE2/E1 and pUC19-λE2/E2sequencing, Sections 3.3.4 and 3.3.5
pUC/M13 RP	caggaaacagctatgac	PUC18/19 +/-	pUC19-λE2/E1 and pUC19-λE2/E2sequencing, Sections 3.3.4 and 3.3.5

Underlined sequence indicates additional nucleotides incorporating a restriction endonuclease target sequence. HSRL = *hsr* locus co-ordinates. +, - denotes priming orientations relative to respective gene orientations.

2.4 Vectors and recombinant plasmids

Vectors and recombinant plasmids constructed and used in this work are summarised in Table 2.4. See Appendix 1 for plasmid maps of constructs made during this study.

Table 2.4. Plasmids and other vectors used in this study.

Plasmid/ vector	Characteristics	Source reference : Appendix 1 reference : Section reference
pHM205	A 2.9 kb HSRL fragment from strain 4298 in pUC18.	(Forester, previous work) : A1.6 : Section 3.6.2
pHM205ΔORF2	The *aphA3* gene inserted into the *Msc* I site of plasmid pHM205	This study : A1.7 : Section 3.6.2
pILL600	Kan^R; pBR322 ori; *aphA3* of *C. coli* plasmid pIP1433 in construct containing pBR322 and *C. coli* pIP1455 sequence	(Labigne *et al.*, 1992): - :Section 3.6.2
pUC18	Ap^R; colE1 ori; blue/white; MCS	(Yanisch-Perron *et al.*, 1985) : A1.5 : Section 3.6.2
pUC19	Ap^R; colE1 ori; blue/white; MCS	(Yanisch-Perron *et al.*, 1985) : A1.2 : Section 3.3.2
pUC19/λE2/E1	A 2.2 kb λE2 fragment (strain 4298 genome derived) in the *Eco* RI site of pUC19	This study : A1.3 : Section 3.3.2
pUC19/λE2/E2	A 1.7 kb λE2 DNA fragment (4298 genome derived) in the *Eco* RI site of pUC19	This study : A1.4 : Section 3.3.2
λE2	A 22 kb genomic fragment from *H. mustelae 4298* inserted into the EMBL3 replacement vector.	(O'Toole *et al.*, 1994) : A1.1 : Section 3.3.1

2.5 Antisera

Table 2.5 lists the antisera used in this study.

Table 2.5. Antisera used in this study.

Name	Function	Source	Relevant characteristics
anti-Hsr (JA6)	primary antibody in Western blots (Section 2.11.2)	P.W. O'Toole, Massey University, Palmerston North, New Zealand	rabbit polyclonal antibody raised against *H. mustelae* 4298 Hsr protein
anti-rabbit IgG – horse radish peroxidase conjugate	secondary antibody in Western blots (Section 2.11.2)	Sigma Aldrich	goat polyclonal antibody raised against rabbit IgG

2.6 DNA preparation

2.6.1 Plasmid preparation

2.6.1.1 Easy plasmid minipreparation (Easyprep)

The Easy Plasmid Minipreparation was performed as described by Berghammer and Auer (Berghammer and Auer, 1993). The method is a crude plasmid preparation, achieved by incubating the cell pellet in a single buffer [10 mM Tris-Cl pH 8.0 / 1 mM EDTA / 15% (w/v) sucrose / 2 mg/ml lysozyme / 0.2 mg/ml Rnase A / 0.1 mg/ml bovine serum albumin (BSA), stored at –20 °C] for lysis and resuspension of the final DNA solution. This procedure was used for simply and quickly screening many recombinant *E. coli* clones after transformation.

2.6.1.2 WIZARD™ plasmid minipreparation

Rapid purification of high purity DNA was carried out using the WIZARD™ Plus Miniprep DNA Purification System (Promega) according to the manufacturer instructions. This method was based on alkaline/SDS lysis (Birnboim and Doly, 1979) of the bacterial cells, followed by selective binding plasmid DNA to a resin, and finally elution with Milli-Q-treated water or TE [10 mM Tris-HCl pH7.5 / 1 mM EDTA]. Typical yields of approximately 10 – 30 µg was obtained from 1.5 – 1.7 mls of an overnight culture.

2.6.1.3 CONCERT™ Rapid Plasmid Purification Miniprep System

An alternative method for rapid purification of pure plasmid from *E. coli* cells was the CONCERT™ minipreparation system (Life Technologies), a modification of the alkaline/SDS lysis procedure (Birnboim and Doly, 1979). Lysis was followed by selective adsorption to a silica-based membrane, washing of DNA and then elution of DNA in a low salt solution i.e. TE or Milli-Q-treated water. The manufacturer instructions were observed, and typical yields of 10- 30 µg of purified plasmid DNA, from 1.5 – 1.7 mls of *E. coli* cells, were obtained.

2.6.2 Preparation of genomic DNA from *Helicobacter* cells

Genomic DNA was isolated from *Helicobacter* cells using the chromosomal quick preparation for *Helicobacter*, a modification of the method of Pitcher *et al.* (1989). Growth from an 18 – 48 hour plated *Helicobacter* culture was harvested with a sterile swab into 1.5 ml of phosphate buffer saline (PBS) [0.1M Na_2HPO_4 / 0.1 M NaH_2PO_4 pH 7.4] and pelleted by centrifugation for 1 min at 20,800 x g, RT. The cell pellet was then resuspended in 100 µl TE [10 mM Tris-HCl pH8.0 / 1 mM EDTA]. Lysis of cells and degradation of RNA was performed by adding 500 µl GES [5 M guanidium thiocyanate / 0.1 M EDTA / 0.5% (w/v) sarkosyl] and mixing together with 2 µl of a 10 mg/ml solution of Rnase and incubating for 10 min at RT. Genomic DNA was precipitated with 1ml of –20°C absolute ethanol, inverting the eppendorf tube and then harvesting by spooling the precipitating DNA onto a pre-made glass hook. The DNA was washed by dipping the DNA into 1ml of –20°C 70% ethanol and then resuspended in 200 µl TE. To digest away proteins associated with the DNA, 20 µl of a 50 mg/ml proteinase K solution was added and incubated at 50°C for 2 - 3 hours. The DNA was recovered from the protease solution by using the precipitation and wash steps as above and then resuspended in 200 - 500 µl for storage at –20°C.

2.6.3 Extraction of bacterial DNA from stomach tissue for PCR analysis

Approximately 30 mg of pelleted ferret stomach tissue, collected as described in Section 2.1.2, was defrosted from –20°C and rinsed twice with 1 ml of sterile PCR grade Milli-Q water by centrifugation at 20,800 x g for 1 min at RT. Subsequently the DNA was extracted as described by the sample processing method of Hammar *et al.* (*Hammar et al.*, 1992). Briefly, the sample was deproteinized by proteinase K digestion followed by phenol-chloroform extraction and finally the DNA was ethanol precipitated (-20°C overnight) and resuspended in 10µl 0.1 x TE [1mM Tris-HCl, 0.1mM EDTA pH 8.0]. DNA precipitation was achieved by addition of 1/10 volumes of 3M Sodium acetate pH 5.6, and then 2.5 volumes of –20°C absolute ethanol. The tube was incubated overnight at –20°C and then pelleted at 20,800 x g for 20 min at room temperature, followed by

washing in 500 µl of –20°C 70% ethanol 20,800 x g, 5 min. The pellet was air dried and resuspended in an appropriate storage solution (see above). Sample aliquots of 50 µl were withdrawn from the procedure at different stages of the preparation and stored at –20°C until required.

2.7 DNA analysis methods

2.7.1 DNA agarose gel electrophoresis

The examination of size, concentration and condition of DNA was performed using horizontal agarose gel electrophoresis following standard methodology (Sambrook *et al.*, 1989). DNA fragments were separated on 0.8 – 1.5% TAE agarose in 1 x TAE buffer [40 mM Tris acetate and 1 mM EDTA] in a Horizon® 58 (minigel) or Horizon® 11.14 gel electrophoresis apparatus (Life Technologies) or a Minicell EC370M (Savant Instruments) at RT. Prior to separation, samples were mixed 1/10 final loading volume of 10 x loading dye [0.5% Bromophenol blue, 0.5% Xylene cyanol, 50% Glycerol, 50 mM EDTA, 1 x TAE]. Running conditions ranged from 50 - 100V for 30 – 120 min RT. Gels were stained with ethidium bromide at approximately 5 µg/ml and then destained briefly in distilled water before examining under UV irradiation on a TMW-20 Transilluminator (Alpha Innotech). Images were captured with an IS-1000 Digital Imaging System (Alpha Innotech). Molecular mass was determined using the 1kb or 1kb plus ladder (Life Technologies) spanning the size range 100bp to 12kb. On occasion lambda DNA was digested with *Hind* III or *Ava* II enzymes for use as size markers using the protocols outlined in Section 2.7.2.

2.7.2 DNA restriction endonuclease treatment

Restriction enzymes were obtained commercially from Boehringer Mannheim (Roche), Life Technologies, and/or New England Biolabs (NEB). Restriction digests were carried out from 1 – 2 hours in their respective buffers according to the manufacturer guidelines. When double digests were required, a buffer to yield optimal enzyme activity for both was used or the supplier's guidelines for double digests were consulted. The enzyme volume in the reaction never exceeded 10% of the total reaction volume

(i.e. ≤5% glycerol). Typically a concentration of 0.2 U/µl of enzyme was maintained for test restrictions. For digestions that were required for subsequent cloning applications, an enzyme concentration of 0.5 – 1.0 U/µl in the reaction was used. Digestion progress was followed by analysis of 1µl of sample on an agarose gel. Enzymes were usually heat-inactivated at 65°C for 20 min once digestion was complete.

2.7.3 DNA quantification

DNA concentration was estimated by comparing the intensity of ethidium bromide stained bands of the linearised sample DNA to known concentration standards after DNA electrophoresis (Section 2.7.1). DNA standards were made by completely linearising 2 µg of commercially obtained pUC18 DNA (NEB) with *Bam* HI enzyme followed by two fold serial dilutions to give DNA concentrations ranging from 5 - 20 ng/µl.

DNA concentration was determined spectrophotometrically at 260nm using a Shimadzu UV-160 Spectrophotometer with quartz cuvettes (Starna) with 1-cm light path. The concentration was calculated on the basis that one A_{260} unit is equal to 50 µg/ml (double stranded DNA) or 33 µg/ml (single stranded NA) taking into account any dilution factor involved in the absorbance reading. The 260/280 absorbance ratio was also calculated to examine the purity of the sample was about 1.8 or above for pure DNA samples with low protein contamination.

2.7.4 Southern blotting and hybridisation

Southern blotting was performed to search for regions amid DNA fragments separated by agarose electrophoresis that were homologous to the probing DNA. The procedure was based on the method of Southern (Southern 1975) and was carried out using the reagents and directions described by the ECL™ direct nucleic labelling and detection systems (Amersham LIFE SCIENCE).

DNA was digested as in Section 2.7.2, subjected to agarose gel electrophoresis, and photo image capture (Section 2.7.1). For size reference marking, 10 ng of *1kb plus*

ladder (Life Technologies) was loaded alongside test samples. The gel treatment, capillary blot assembly and disassembly was done according to the ECL manual (Amersham). Overnight transfer of DNA to positively charged Hybond-N+ membrane (Amersham) was facilitated using 20 x SSC buffer transfer medium. After the completion of the transfer, the marker lanes were cut away from the rest of the sample for separate processing. The transferred DNA was covalently fixed to the nylon filter by exposure to UV light for 30s on the blotted side of the membrane using a TMW-20 transilluminator (Alpha Innotech).

The principle of labelling and detection of the ECL kit is based on enhanced chemiluminescence. The denatured single stranded probe was labelled by forming covalent crosslinks with a positively charged polymer, which is complexed with horseradish peroxidase. Once the probe has hybridised to immobilised target DNA and the excess unbound probe was removed, the enzymatic action of the horseradish peroxidase is coupled to a light producing reaction, which produces blue light that gave a signal on an autoradiograph.

Probe DNA was generated by PCR amplification (Section 2.8) followed by purification from an agarose gel (Section 2.9.1.1). The sample was then concentrated to 10 ng/µl by evaporation of water in a Speedvac (Savant Corporation). The probe DNA for the marker lane was 100 ng of *1kb plus* ladder in 10 µl final volume. The probe was freshly prepared each time. Unlabelled probe DNA was stored at –20°C. Hybridisation and washes were carried out at 42°C as per supplier recommendations. Instructions were strictly observed for conducting the detection reaction. The length of exposure to X-ray film (Fuji) that gave the best signals was approximately 30s before developing for 2 min in each of the Kodak developer and fixer with a brief tap water rinse in between. Finally the autoradiographs were washed under running water and then air dried at 37°C.

2.7.5 DNA sequencing

DNA sequence determination of plasmid or PCR templates was performed at the Massey University DNA sequencing Facility (Massey University, Palmerston North,

New Zealand) on an ABI Prism 377 DNA sequencer (Applied Biosystems). Sequencing reactions were performed based on the method of Sanger *et al.* (Sanger *et al.*, 1974) with the resulting sequence data received in the form of chromatograms that could be viewed and edited using the ABI Prism software (Editview).

Plasmid and PCR product DNA templates were prepared at concentrations of 200 ng/µl and 20 ng/µl, respectively, using techniques described elsewhere in this chapter. Sequencing primers were at a concentration of 0.8 pmol concentration. Edited sequence data determined as above, was analysed and manipulated using the Geneworks package (IntelliGenetics), the GeneJockey programme (Biosoft) and DNA Strider (Dr Christian Marck, France). Sequences were also submitted to database searches for related sequences by accessing the National Centre for Biotechnology Information (NCBI) and using the basic local alignment search tool (BLAST) algorithm (Altschul *et al.*, 1990). Contiguous sequence was assembled using the Geneworks system (IntelliGenetics). Multiple sequence alignments were created using the Clustal V programme (Higgins *et al.*, 1992) using default parameters.

2.8 DNA amplification by polymerase chain reaction (PCR)

DNA segments were amplified by the polymerase chain reaction; a primer-directed enzymatic process that enables the amplification of DNA between two flanking oligonucleotide primers.

Reaction volumes varied from 10 – 100 µl, depending on the amount of DNA required for subsequent applications. Reactions were carried out in thin walled 0.2 ml tubes, in a thermocycling instrument (FTS-960 Microplate thermal sequencer, Corbett Research or Omni-E cycler, Hybaid), and subjected to 25 - 35 cycles of denaturation, annealing, and extension after an initial extended denaturation step of 1 - 2 min. A final extension step that was four times longer than the cycling extension was performed to ensure complete synthesis of the DNA products. PCR products were stored at –20°C.

The thermocycling conditions for each reaction were influenced by a number of factors. The manufacturer guidelines for each polymerase were consulted for suitable

denaturation (92°C or 94°C) and extension (68°C or 72°C) temperatures. The annealing temperatures were determined by calculating the melting temperature of each primer used for amplification and then setting the annealing temperature two degrees (°C) below the lower melting temperature of the two primers. The melting temperatures were estimated using the formula T_m=2 x n(A + T) + 4 x n(G + C). The times for each step and number of cycles were adjusted according to the polymerase manufacturer guidelines. The extension times were dependent on the length of the expected DNA fragment.

A typical reaction included the following components: 200 µM deoxynucleoside triphosphate mix; 1.5 - 2.5 mM $MgCl_2$ (for *Taq* polymerase and eLONGase) or $MgSO_4$ (for *Pwo* polymerase); 0.2 µM of each primer; 1 x manufacturers PCR buffer(s); 0.025 - 0.05 U/µl of recombinant *Taq* DNA polymerase (Life Technologies) or *Pwo* Polymerase (Roche) which generated high fidelity blunt ended fragments for subcloning or 1 or 2 µl eLONGase enzyme mix (Life Technologies) for amplification of products >4 kb; template DNA (genomic 50 - 100 ng or plasmid 0.5 - 2 ng); and sterile Milli-Q water to make up the rest of the reaction volume. A master mix of common components of the reaction was routinely used to simplify PCR set up. Appropriate positive (when possible) and negative control reactions were processed alongside sample PCR reactions.

In addition, PCR screening of transformant clones was sometimes carried out using bacterial cells as PCR templates. For *E. coli* and *H. mustelae* colonies, the entire colony was lifted using a sterile disposable loop and suspended in 50 µl of LB and BHI, respectively (Section 2.2). *E. coli* suspensions were used directly at 1/10 the final PCR reaction volume, while the *H. mustelae* colony suspensions were heated to 95°C for 5 minutes and then cooled briefly at room temperature prior to addition to the reaction mix.

2.9 Cloning strategies

DNA cloning was achieved by: generating DNA fragments with compatible ends; joining DNA fragments in a ligation reaction; transformation of recombinant DNA into

suitable host cells; selecting for clones containing the recombinant DNA using standard genetic techniques (Sambrook *et al.*, 1989).

2.9.1 DNA preparation

2.9.1.1 Vector DNA

Plasmid vector DNA was usually prepared using either WIZARD™ (Section 2.6.1.2) or CONCERT™ (Section 2.6.1.3). The resulting DNA was then subjected to restriction endonuclease digestion as described in Section 2.7.2 followed by fragment separation by agarose gel electrophoresis (Section 2.7.1). Subsequent removal of the band from the gel was performed using a clean sterile scalpel and the visual guidance of a UVGL-58 Mineralight lamp (UV Products Ltd) set on long wave UV setting to reduce damage to the DNA during the gel excision process. DNA was recovered from agarose using the High Pure™ PCR product purification kit (Roche) and the method for purifying DNA from a 100mg slice of agarose gel in Nucleic acid isolation and purification handbook (Boehringer Mannheim). DNA fragments were concentrated, if necessary, either by evaporation (Section 2.7.4) or ethanol precipitation (Section 2.6.3) and then the concentration was estimated as in Section 2.7.3. For PCR generated vectors with incorporated restriction sites, the DNA was first purified using the High Pure™ PCR product purification kit (Roche) according to the manufacturers instructions, prior to the restriction digest step above. If required, vectors were treated to remove the 5' phosphate groups from the DNA in order to prevent self-ligation using one unit of calf intestinal alkaline phosphatase, CIP, (NEB) per pmol of DNA ends. Dephosphorylation was performed as per manufacturer guidelines, but the vector DNA was passed through a High Pure™ PCR product purification column to recover the vector DNA, instead of phenol extraction and alcohol precipitation.

2.9.1.2 Insert DNA

DNA to be ligated to vector DNA was produced by digestion of plasmid/ phagemid DNA by restriction endonuclease treatment, followed by purification from an agarose gel (refer to Section 2.9.1.1), prior to ligation.

2.9.2 Ligation

Ligations were performed using 0.2 – 0.5 U T4 DNA ligase (Roche) in a 10 µl reaction volume to join together ligation-compatible termini of insert DNA and vector DNA. Accepted molar ratios of 1 – 5 to 1 of insert to vector DNA were observed for reactions, depending on the yield of gel-purified fragments generated as in Section 2.9.1.1. Approximately 10 – 50 ng of plasmid DNA was used for each ligation experiment. Master reaction mixes, which were employed for all ligations, contained all of the ligation reagents excluding insert DNA molecules. These master mixes facilitated the set up of vector-only control ligation reactions in parallel with sample ligations. Ligations of cohesive termini were achieved by incubation overnight at 4°C, whilst ligation of blunt ends was performed at 16°C.

2.9.3 Transformation

2.9.3.1 Preparation of competent bacterial cells

A quick method of making *E. coli* strain ER2206 competent cells for same day use was acquired from guide to the GST gene fusion system (*3rd Ed*, Pharmacia Biotech), which was based on that of Chung and colleagues (Chung *et al.*, 1989). A single colony from an overnight LB plate containing 10 µg/ml tetracycline was innoculated into 50 – 100 ml of LB broth and grown to an absorbance reading at 600 nm (A_{600}) of 0.4 - 0.5 as set out in Section 2.1. Cells were harvested by centrifugation at 2,500 x g for 15 mins at 4°C, and then resuspended in 1/10 the original culture volume of ice cold TSS buffer [1% tryptone, 0.5% yeast extract, 0.5% NaCl, 10% PEG 3350, 5% Dimethylsulfoxide (DMSO) and 20mM $MgCl_2$ pH6.5] and kept on ice for use.

Frozen competent *E. coli* cells were prepared using a modified method of Hanahan (Hanahan, 1983). 100 – 500 ml of LB (Section 2.2) was inoculated with a 1/50 volumes of an overnight culture of *E. coli* strain ER2206 cells and agitated at 37°C until the A_{600} was within the range of 0.4 – 0.6 in a spectrophotometer (Science and Technology). Cells were then harvested by centrifugation at 5,000 x g for 5 min, at 4°C. The cell

pellet was resuspended in 1/5 the original culture volume of buffer CM1 [10 mM sodium acetate pH 5.6/ 5 mM NaCl/ 50 mM $MnCl_2$] and incubated on ice for 20 min before repeating centrifugation as above. After carefully removing supernatant, the cell pellet was resuspended in 1/50 of the original cell volume of sterile ice cold buffer, CM2 [10 mM sodium acetate pH 5.6/ 50 mM $MnCl_2$/ 5% glycerol/ 70 mM $CaCl_2$]. Cells were stored immediately at –70°C in aliquots of 50 - 100 µl.

To test the competence of the cells made by the above methods, 100 µl of competent cells was mixed with 50 pg of pUC19 obtained commercially at 5 pg/µl (Life Technologies). A transformation reaction was performed as in Section 2.9.3.2, and transformants carrying the pUC19 plasmid were selected for by plating on LBA containing 200 µg/ml ampicillin. Transformation frequencies were calculated as the number of colony forming units per microgram (cfu/µg) of DNA including the dilution factor of the cells during the transformation process. Transformation frequencies of $f10^{-6}$ cfu/µg were deemed acceptable for test transformations.

2.9.3.2 Transformation of E. coli

Transformation of *E. coli* cells was performed based on the standard methodology of Sambrook *et al.* (Sambrook *et al.*, 1989). Cells were kept or defrosted on ice, depending on the method of preparation. 50 – 100 µl of competent cells were mixed with 10 µl final volume of ligation mix incorporating undiluted ligation reaction (0 - 10µl), containing approximately 10 ng DNA made as in Section 2.9.2, the volume made up to 10 µl with sterile Milli-Q water. The transformation mix was incubated 45 min on ice and then subjected to heat shock at 42°C for 1 min followed by a brief 2 min incubation on ice. The transformed cells were allowed to recover by gentle agitation at 37°C for 60 - 90 min diluted in 9 volumes of Luria Broth (Section 2.2). 100 µl of recovered cells was spread onto LBA plates containing the relevant selective reagents (Section 2.2) and incubated overnight at 37°C. Appropriate positive control transformations were included frequently, while the water only and vector only control transformations were always processed.

2.9.3.3 Transformation of H. mustelae

The natural competence of the *Helicobacter* species was exploited using a method based on that outlined by O'Toole *et al.* (O'Toole *et al.*, 1994b). One to two day old cell lawns were harvested from solid media support into BHI broth (Section 2.2) and adjusted to an A_{600} of 0.1 at using a visible spectrophotometer (Science and Technology). 100 µl of these cells was spread onto a CBA plated and grown for approximately 16 hours in microaerobic conditions (Section 2.1). One plate was harvested into 1 ml of CO_2-saturated BHI and the OD_{600} was adjusted to 1.0. The transformation reaction was set up using 100 µl of harvested cells and 10 µl of recombinant plasmid (2 - 5 µg) made to volume with sterile Milli-Q water. The transformation was incubated at 37°C in a CO_2 incubator (Revco Scientific) in loosely capped sterile cryovials (Sarstedt) for 4 hours with intermittent mixing. Selection for transformants was achieved by spreading 100 µl of suitably diluted (10^0 – 10^{10} fold) cells over appropriate selective CBA plates (Section 2.2) and incubating at 37°C. First cultures were grown for 2 days in an anaerobic jar containing a CampyGen sachet (Oxoid) before transferring to a CO_2 incubator for the remaining 2 – 7 days. Alternatively 25 µl of the appropriately diluted transformed cells were spread onto quadrants of the CBA based plates before incubating as above. For colony counting 3 - 4 quadrants were spotted with 25 µl of each sample to facilitate the calculation of the tranformation frequencies. The remaining cells were usually collected by a brief 15s centrifugation at 20,800 x g, at RT, resuspended in 1 x BHI/20% glycerol and stored at –80°C as a back up. Transformant colonies usually appeared after 3 - 4 days incubation at ratios of approximately 10^{-5} - 10^{-6} (calculated by the number of transformed colonies divided by the number of surviving *H. mustelae* cells).

2.10 Protein sample preparation

H. mustelae cultures grown on CBA plates were harvested with a sterile swab into 1 ml of 1 x PBS [0.1M Na_2HPO_4 / 0.1 M NaH_2PO_4 pH 7.4] taking care not to carry over any agar from the plate. To standardise the amount of cells, the optical density at wavelength 600 nm (A_{600}) was adjusted to 1.0. Alternatively, the cell pellet wet weight was determined as follows. The cell suspension was subjected to centrifugation at

20,800 x g for 1 min at RT. All of the suspension liquid was removed and discarded and the nett cell weight determined in a preweighed eppendorf tube. The cell pellet was finally resuspended in an appropriate amount of 1 x PBS to a concentration of 10 - 50 mg/ml. Samples were boiled for 5 min and stored at –20°C.

2.11 Protein analyses

2.11.1 Protein electrophoresis

Proteins were separated and analysed on vertical, SDS denaturing, polyacrylamide gels using the Mini-Protean® II Electrophoresis cell system (Bio-Rad), as per the manufacturers instructions. The gel set up was based on that of Laemmli (Laemmli, U.K., 1970), a discontinuous (two-layer) gel with all solutions made according to the "Reagents and Gel preparation for SDS-PAGE slab gel (Laemmli buffer system)" in the manufacturer instruction manual. The base resolving gel consisted of either 7.5% acrylamide whilst the upper stacking gel contained 4% acrylamide. After assembly of the SDS PAGE gels, the wells were rinsed and equilibrated for 5 min in 1 x electrode (running) buffer [0.3% (w/v) Tris base, 1.45% (w/v) glycine, (w/v) 0.1% SDS pH 8.3] prior to applying current. During this time protein samples were prepared for loading into wells by boiling for 5 min in the presence of a SDS reducing buffer [62.5mM Tris-Cl pH 6.8, 10% (v/v) glycerol, 2% (w/v) SDS, 0.01% bromophenol blue, 5% (v/v) 2-mercaptoethanol]. A ratio of sample to sample buffer of 1:4 was used. Prestained protein molecular weight standards (Life Technologies), spanning the range 14 to 200 kDa, were included on all gels (2 – 5 µg per lane). A constant current of 30 mA was applied for 30 – 45 min to each gel, followed by gel disassembly, and staining by soaking in a Coomassie Brilliant Blue solution [40% (v/v) methanol, 10% acetic acid, 0.25% (w/v) Coomassie Brilliant Blue (Sigma) for 20 – 60 min. Proteins became visible after washing several times in destaining solution [40% (v/v) methanol, 10% acetic acid]. Stained gels were dried using a vacuum gel dried (BioRad Laboratories, California, U. S. A.) according the manufacturers instructions and stored at RT.

2.11.2Western blotting and immunodetection

The Western blot system used was based on the method of Towbin and colleagues (Towbin *et al.*, 1979), whereby proteins separated by gel electrophoresis are transferred to nitrocellulose membrane (Satorius) and then specifically detected using indirect immunodetection of specific antigens via a secondary enzyme reaction involving horseradish peroxidase. Proteins were separated as per Section 2.11.1 and the resulting gels were pre-equilibrated in transfer buffer [25mM Tris pH 8.3/ 192 mM glycine/ 20% methanol] for 5 min prior to the transfer of separated proteins to nitrocellulose paper. The protein transfer was achieved using the Mini Trans-Blot® Electrophoretic Transfer Cell (BioRad) system. Unit assembly and transfer were performed as instructed by the manufacturers technical manual. The transfer was carried out for 90 min at 40 volts in transfer buffer and then disassembled and the blots were rinsed and equilibrated in 50 ml of TBS [10mM Tris-HCl pH 7.5, 0.9% (w/v) NaCl pH7.4]. To block non-specific binding sites, on the surface of the nitrocellulose, an incubation was performed for at least 20 min in approximately 20 ml of SM-TBS [1% (w/v) skim milk powder in TBS]. The SM-TBS was then replaced with 10ml of an appropriately in freshly diluted primary antibody (typically 1/5000 fold dilution in SM-TBS). The primary antibody incubation time was approximately 60 min at RT with gentle agitation. The unbound primary antibody was removed by three five-minute washes, the first wash in TBS, the second in TBS containing 0.05% (w/v) Tween 20, and the third wash in TBS. Secondary antibody incubation was performed for 1 hr at RT with an antibody, appropriately diluted (1/1000) in SM-TBS with gentle agitation on a shaking platform. Repeating the washing steps above washed off the unbound secondary antibody. The secondary antibody used in this study was anti-rabbit IgG conjugated to horseradish peroxidase enzyme (Serotec) (Section 2.5). The washed blot was soaked in detection solution and developed in the dark until visible signals were sufficiently observed. The detection solution was made by dissolving 30mg of 4-chloro-1-napthol (Sigma) in ice cold methanol (kept at –20°C), mixing this with 50 ml of TBS and finally adding 50µl of 30% stabilised hydrogen peroxide (BDH) to complete the solution. Several rinses of Milli-Q water stopped the enzymatic reaction and the blots were then stored dry at RT.

2.12 Microscopy

The visualisation of bacterial cells and other cell debris was achieved using a phase contrast transmitted-light microscope KM (Carl Zeiss) at 400 times magnification.

3.0 RESULTS

3.1 Basic analysis of the 12 kb *hsr* locus (HSRL)

Prior to the commencement of this study, DNA sequences significantly related to the *hsr* gene of strain 4298 (accession L15629) were detected upstream of the *hsr* gene by BLAST nucleotide homology searches of the recently sequenced DNA (Section 1.3.3). The BLAST search located the boundaries of this upstream *hsr*-like sequence and an *hsr* locus of 12144 bp, containing the *hsr* gene and the upstream repeats, was thus defined.

A detailed analysis of DNA sequences that were repeated in the entire 12144 bp of the HSRL was performed using the DNA Strider software package (Dr Christian Marck, France). The search parameters used were a minimum repeat length of 16 nucleotides with 100% identity. Imperfect repeats together complexed the search such that the memory capacity of the program could not cope with the number of alignments. Direct repeats matching the above search parameters were not found within the *hsr* gene. To find inverted repeats, HSRL coordinates-7482 – 12144 were inverted and scrutinised for repeated sequences. In total, greater than 200 pairwise DNA alignments were generated, containing 115 sequences that were repeated 2 – 5 times and up to 741 nucleotides in length. The repeats were categorised into groups and subgroups according to their location within the HSRL. Upstream repeat group (USRs) sequences were located upstream of the *hsr* gene and *hsr* repeat group comprised repeats in the *hsr* gene. Repeat groups (USR or HSR) were further classified into subgroups by the following criteria: Subgroup U repeats consisted of sequences found only upstream of the *hsr* gene (HSRL coordinates 1-7481), while subtype H sequences were found only in the *hsr* gene. Subgroup HU repeats were located both upstream and within the *hsr* gene. No H (*hsr* only) repeats were observed with these search parameters. Table 3.1 summarises the frequency and types of repeats observed and the complete repeat details are listed in Appendix 3 (refer to Section 3.4). Figure 3.1 illustrates the area of the HSRL covered by repetitive DNA and demonstrated that the *hsr* gene repeats were confined to the first half of the *hsr* gene (HSRL, 7485 - 9847 and Accession L15629, 125 - 2490). Consequently, this area was designated the repeat region of the *hsr* gene.

Table 3.1 Summary of the repeat sequence frequency.

Number of times repeated	No. of repeats	Length range (bp)	Frequency of repeat subtype H(U)	Frequency of repeat subtype U
2	59	16 - 741	27	32
3	41	16 - 137	21	20
4	13	16 - 30	6	7
5	2	18 - 34	2	0
Total	**115**	**16 - 741**	**56**	**59**

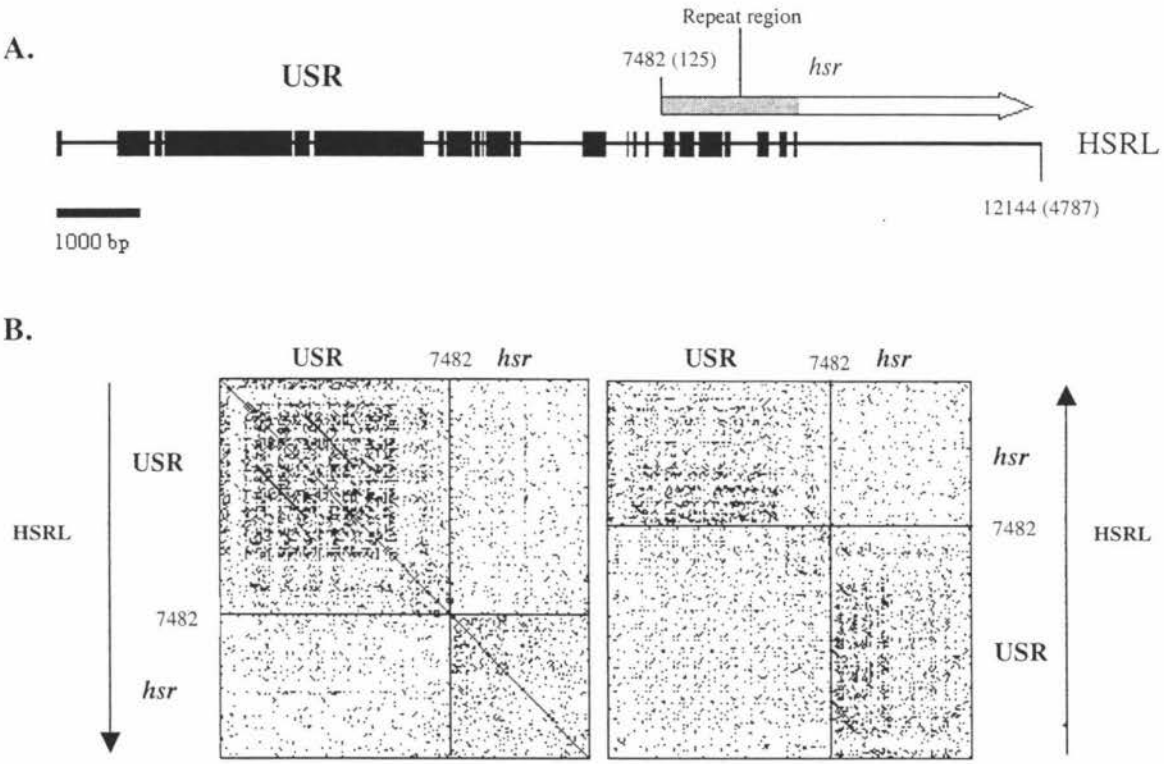


Figure 3.1 **Occurrence of repetitive DNA sequences in the 12 kb *hsr* locus (HSRL).**

A. Diagram showing the repeat sequence frequency within the 12144 bp HSRL based on the results of the DNA Strider repeats search. All repeat sequences are shown as thick black boxes along the main HSRL line. Labelled arrow indicates the length of the *hsr* gene and orientation with respect to the HSRL. Grey area inside *hsr* arrow depicts the repeat region of the *hsr* gene. Position markers point out the HSRL co-ordinates next to the accession L15629 co-ordinates in closed brackets. The black bar represents 1000 bp scale. **B.** Dot matrices show the repetitive nature of the *hsr* locus. The long black vertical arrows indicate the orientation of the HSRL in the alignment and the individual sections of the HSRL are labelled. Lines inside the matrices mark the border of each labelled section. Reference HSRL co-ordinates are shown for the start of the *hsr* gene. USR = upstream repeats group.

3.2 Sequence variation of the *hsr* gene in *H. mustelae*

Repetitive sequence was apparently restricted to the first half of the *hsr* gene, i.e. the repeat region. The existence of large numbers of DNA sequence repeats in the upstream region suggested that the DNA sequence in the *hsr* repeat region could be potentially replaced with DNA from the upstream, thus generating variability in the *hsr* repeat region. To investigate this further DNA sequences of PCR products amplified from the *hsr* repeat region of different *H. mustelae* isolates, collected from New Zealand and American origin ferrets, (Section 1.2.2) were compared. The PCR products were partially sequenced from one end of the molecule and the resulting sequences were aligned.

3.2.1 *Helicobacter* isolation from stomachs of New Zealand ferrets

To investigate the possibility of variability of the repeat region of the *hsr* gene, a collection of strains of *Helicobacter mustelae* was needed. There was only one strain of *H. mustelae* (strain 4298) available to work with in New Zealand. It was unknown whether *H. mustelae* colonised New Zealand ferrets. Consequently, an attempt was made to isolate *H. mustelae* from the stomachs of ferrets sacrificed as part as an unrelated study to determine if *H. mustelae* was present in the New Zealand ferret population. Stomachs of 21 ferrets were obtained from two separate sources in the North Island of New Zealand. Four ferrets were from a breeding colony in Hamilton, and 17 ferrets were gathered from the lower Hawkes Bay (Section 2.1.2 for details). The time from stomach collection until processing ranged from an hour to at least three days depending on how long before traps were checked. For culture, stomachs were processed as described in Section 2.1.2. Cells matching *Helicobacter* by colony morphology and microscopic characteristics (Section 2.12) were additionally tested by standard microbiological techniques, (urease and oxidase activity) as described in Section 2.1.2.1. Table 3.2. summarises stomach processing details and success of culture. For PCR detection of *Helicobacter*, genomic DNA was extracted from cultured *Helicobacter* cells as described in Section 2.6.2 and also from thawed tissue sections as described in Section 2.6.3. PCR analysis was subsequently performed on culture positive samples according to the guidelines set out in Section 2.8 using primers

HS16sF and HS16sR (Table 2.3), designed to specifically amplify part (734 bp) of the 16s ribosomal gene of *Helicobacter* strains.

Six out of 21 ferret stomachs yielded viable Gram negative, oxidase positive, and urease positive rods. In three out of four colony-bred ferret stomachs (F6, F7, and F21), and three out of seventeen wild ferret stomachs (F8, F11, and F15), isolates were confirmed as *Helicobacter* by PCR analysis.

Table 3.2 Ferret stomach processing details

Ferret ID (F#)	Location	Stomach condition[a]	Culture growth
1	Dannevirke	0do,P	-
2	Dannevirke	0do,P	-
3	Dannevirke	3do,P	-
4	Dannevirke	1do,G	-
5	Dannevirke	1do,G	-
6	Hamilton	0do,VG	+
7	Hamilton	0do,VG	+
8	Dannevirke	1do, G	+
9	Dannevirke	1do, G	-
10	Dannevirke	1do, G	-
11	Dannevirke	1do, G	+
12	Dannevirke	1do, F	-
13	Hamilton	0do, VG	-
14	Dannevirke	3do, P	-
15	Dannevirke	1do, G	+
16	Dannevirke	1do, G	-
17	Dannevirke	3do, P	-
18	Dannevirke	3do, P	-
19	Dannevirke	1do, F	-
20	Dannevirke	1do, F	-
21	Hamilton	0do, VG	+

Stomach condition details[a] do (days old), specifies the number of days between collecting dead ferret from trap and processing as in Section 2.1.2. Condition of ferret stomachs range from poor (P), fair (F), good (G) to very good (VG).

3.2.2 Molecular analysis of the New Zealand (N. Z.) *H. mustelae* isolates

3.2.2.1 16S ribosomal gene analysis of Helicobacter strains isolated from N.Z. ferrets

For confirmation of *Helicobacter* species, DNA sequence of a representative portion of the 16S ribosomal (rRNA) gene containing variable stretches of DNA, which would distinguish between different species of Helicobacter, was determined. Genomic DNA was isolated from each New Zealand ferret *Helicobacter* isolate and laboratory strain

4298 (Section 2.6.2), and PCR was performed, employing oligonucleotides HS16SF and HS16SR (Table 2.3) designed to specifically amplify 16S sequence of the *Helicobacter* genus, as outlined in Section 2.8. The resulting, gel purified (Section 2.9.1.1), 735 bp segment of the 16S rRNA gene from each strain was subsequently sequenced using the forward amplification primer, HS16SF (Section 2.7.5). The DNA sequences obtained were edited using Editview (Section 2.7.5) and then subjected to a NCBI BLAST nucleotide database search. The *Helicobacter mustelae* strain 4298 16S rRNA gene sequence was submitted as a positive control DNA for comparision and alignment reference.

In all seven strains analysed, the highest identity match was observed with the 16S ribosomal rRNA gene of *H. mustelae* strain 91-292-E1A (Accession M88156, Fox *et al.*, 1992), *H. mustelae* strain ATCC 43772 (Accession M35048, Paster *et al.*, 1991), and *H. suncus* strain Kaz-2 (Accession AB006148) in that order. The full alignment of all 10 DNA sequences is shown in Figure 3.2 and the alignment results are summarised in Table 3.3.

Table 3.3 Summary of 16S DNA alignment results for the *Helicobacter* isolated from ferrets from two separate regions of New Zealand.

Strain/ accession	Source	Length (bp) of aligned sequence	Number of mismatches	# Ns in sequence	% identity to Strain 4298
4298	Ferret	696	-	-	100.0
F6	Ferret	519	-	-	100.0
F7	Ferret	694	12	-	98.3
F8	Ferret	245	-	-	100.0
F11	Ferret	512	-	-	100.0
F15	Ferret	338	-	-	100.0
F21	Ferret	554	4	5	98.4 - 99.3
91-292-E1A /M88156	Ferret	696	8	3	98.9
43772/ M35048	Ferret	696	8	10 (3*)	97.4 (98.4*) – 98.8
Kaz-2/ AB006148	House musk shrew	696	16		97.8

*Alternative lower identity range calculation to accommodate suspected poor sequencing resolution.

The number of nucleotides used in the alignment was dependent on the quality of the sequence obtained from the PCR fragments. The percentage identity ranged from 98.3-100% over a range of 242 – 696 bp of single stranded DNA sequence. The percentage identity was calculated as the aligned length of sequence of query DNA (bp), less the number of mismatches to strain 4298 (bp), all divided by the length of the sequence of

Figure 3.2 Multiple alignment of partial 16S DNA sequences from *Helicobacter mustelae* strains isolated from New Zealand ferret stomachs.

Strain names are annotated to the left of the aligned sequences. Strain 4298 is the laboratory reference strain for *H. mustelae* species. F6, F7, and F21 are *Helicobacters* isolated from New Zealand breeding colony ferrets while F8, F11 and F15 are *Helicobacter* isolates from wild New Zealand ferrets. The sequences of accessions M88156, M35048, and AB006148 were identified by a NCBI BLAST nucleotide homology search as the three best identity matches to the *Helicobacter* strains in question. Differences to the 4298 16S DNA sequence are displayed and in lower case. Gaps in the sequence alignments are marked with a dash (-). Nucleotides exactly matching strain 4298 are not shown. The end of each DNA sequence in the alignment is denoted with an E.

```
                    10         20         30         40         50         60         70
Hm4298      CACTGGAACT GCAGACACGG TCCAGACTCC TACGGGAGGC AGCAGTAGGG AATATTGCTC AATGGGCGAA
M88156                 -
M35048                 -
AB006148               -
F7                                                                                  g
F21                                                                                 n

                    80         90         100        110        120        130        140
Hm4298      AGCGTGAAGC AGCAACGCCG CGTGGAGGAT GAAGGTTTTA GGATTGTAAA CTCCTTTTCT AAGAGAAGAT
M88156        c
M35048        c
AB006148      c
F7          c                                                           g  t        a
F21         n                                                           n  n        a

                   150        160        170        180        190        200        210
Hm4298      AATGAAGGTA TCTTAGGAAT AAGCACCGGC TAACTCCGTG CCAGCAGCCC GCGGTAATAC GGAGGGTGCA
M88156         c                                                  -
M35048         c                        nn                              n n            n
AB006148    t  c                                                  -
F7           c           a c
F21          c           n c

                   220        230        240        250        260        270        280
Hm4298      AGCGTTACTC GGAATCACTG GGCGTAAAGA GCGCGTAGGC GGAGTAATAA GTCAGATGTG AAATCCTGTA
M35048                            n
AB006148                                                 g
F21                                                                                 g
F8                                          E
```

Figure 3.2 Multiple alignment of partial 16S DNA sequences from *Helicobacter mustelae* strains isolated from New Zealand ferret stomachs. (continued...)

```
                290        300        310        320        330        340        350
Hm4298     GCTTAACTAC AGAACTGCAT TTGAAACTGT TATTCTAGAG TGTGGGAGAG GTAGGTGGAA TTCTTGGTGT
AB006148                                                a
F15                                                                 E

                360        370        380        390        400        410        420
Hm4298     AGGGGTAAAA TCCGTAGAGA TCAAGAGGAA TACTCATTGC GAAGGCGACC TACTGGAACA TTACTGACGC
M35048          n

                430        440        450        460        470        480        490
Hm4298     TGATGCGCGA AAGCGTGGGG AGCAAACAGG ATTAGATACC CTGGTCGTCC ACGCCCTAAA CGATGAATGC
M88156                                                  a
M35048                                                  a
AB006148                                                a

                500        510        520        530        540        550        560
Hm4298     TAGTTGTTGG GGTGCTTGTC ACTCCAGTAA TGCAGTTAAC ACATTAAGCA TTCCGCCTGG GGAGTACCGG
M88156                                                                              -
M35048                                                                              -
AB006148                                                                            -
F21                                                                          E
F6                                      E
F11                           E
                                                                                    .

                570        580        590        600        610        620        630
Hm4298     TCGCAAGATT AAAACTCAAA GGAATAGACG GGGACCCGCA CAAGCCGGGG AGCATGTGGT TTAATTCGAT
M88156                                                  g t                         n
M35048                                                  g t                         n
AB006148                                                g t                         a
F7                                                      gt

                640        650        660        670        680        690        700
Hm4298     GCTACGCGAA GAACCTTACC TAGGCTTGAC ATTGATAGAA TCTGCTAGAA ATAGCGGAGT GTCTAGC
M88156     nn                                                                      E
M35048     nn                                                                      E
AB006148    a                                          a     g     t           E
F7         t                          g                                    E
```

query DNA. Where ambiguous (N) residues were taken into account, the resulting ranges of identity values are shown in Table 3.3. It is worth noting that seven out of 10 Ns included in the calculations for M35048 were most likely the result of poor sequencing resolution as determined by examination of the multiple alignment in Figure 3.2. An alternative percentage identity value, shown in Table 3.3, was calculated to take this factor in to consideration. These results suggested that the ferret *Helicobacter* isolates were most likely to be *Helicobacter mustelae.*

3.2.2.2 Total cellular protein profiles and Western blot analysis of H. mustelae isolates obtained from N.Z. ferrets

A two-day-old culture, grown on CBA, of each of the *Helicobacter* isolates and strain 4298 was harvested into 1 x PBS, standardised by wet weight and prepared for protein electrophoresis as described in Sections 2.10. Separated proteins (Section 2.11.1) were subsequently transferred onto nitrocellulose membranes, and detected using polyclonal rabbit antibody against the Hsr protein, JA6 (O'Toole et al., 1994) and anti-rabbit IgG conjugated to horseradish peroxidase as described in Section 2.11.2.

The total protein profiles for each of the strains, detected by coomassie blue staining, were very similar to that of the control strain 4298. All extracts contained a dominant ~150 kDa band, the presumptive Hsr protein, that could be visualised easily in the coomassie-stained polyacrylamide gel (Fig. 3.3a). The size appeared to differ slightly between isolates, F15 appearing largest. Western blot analysis, using anti-Hsr antibody (JA6, O'Toole *et al.*, 1994), confirmed the identity of the 150 kda protein band as Hsr (Fig. 3.3b). The additional bands detected on the immunoblot were most likely breakdown products of the Hsr protein. These results show that the strains isolated from New Zealand ferrets were *H. mustelae* and that all strains expressed the Hsr protein at levels comparable to strain 4298. However, strain-to-strain differences were observed in the reactivity of Hsr to anti-Hsr antiserum. In order from the highest to lowest reactivity to Hsr anti-serum was: 4298>F11>F7>F8>F15>F6/F21.

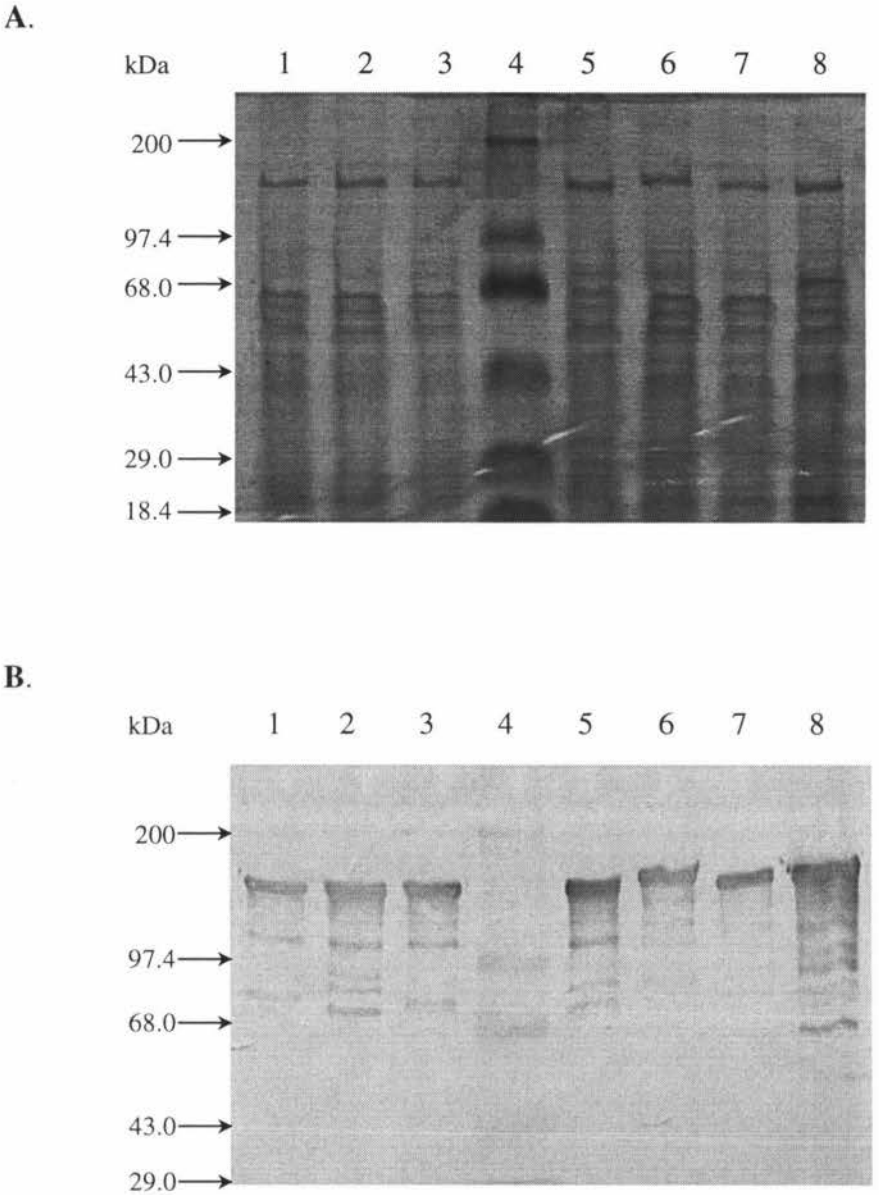


Figure 3.3 Hsr is present in New Zealand ferret *H. mustelae* isolates.

SDS-PAGE and Western immunoblot analysis of *H. mustelae* cell lysates using polyclonal antisera to Hsr. Proteins were separated on duplicate denaturing 7.5% polyacrylamide gels, and subsequently Coomassie stained (panel A) or blotted (panel B) and the hsr protein detected with the anti-Hsr (strain 4298) serum, JA6 (O'Toole *et al.*, 1994).

Lanes: 1, F6; 2, F7; 3, F8; 4, Protein size marker (Life Technologies); 5, F11; 6, F15; 7, F21; 8, *H. mustelae* strain 4298.

3.2.3 Southern blot analysis to investigate organisation and conservation of the HSRL of seven *H. mustelae* isolates

Southern blot analysis was employed to examine the extent of conservation of the *hsr* locus (HSRL) among *H. mustelae* strains before initiating an expensive sequencing project to compare corresponding nucleotides. *Helicobacter* strain 4298 was used as a reference locus since the DNA sequence was known. The probe was designed to hybridise to all DNA fragments containing repetitive sequence as described in Section 3.1.

For Southern blot analysis, 2 µg of genomic DNA of *H. mustelae* strains F6, F7, F8, F11, F15, F21, and 4298 was digested with a *Dra* I and a *Pvu* II/ *Cla* I (Section 2.7.2). Progress of the restriction enzyme digest was assessed by DNA gel electrophoresis of 200 ng of each sample (Section 2.7.1). Digestion was deemed to be acceptably complete when a uniform smear of DNA fragments along the length of agarose gel was visible. The density and distribution of DNA fragments was characteristic of restriction enzymes with a six base pair target sequence. A further 500 ng of each digested sample was then loaded onto a 1% TAE agarose gel for Southern blot analysis (Section 2.7.4). The restriction analysis programme Gene Jockey (Biosoft) was employed to determine the digestion pattern of the strain 4298 HSRL with several restriction enzymes known to cleave *Helicobacter* genomic DNA. The sequence of the repeat region was subjected to the dot matrix program of Geneworks 2.5 (Intelligenetics) to create a matrix pattern to find a highly repetitive region suitable for probe amplification by PCR. By linking this information with the expected restriction pattern, the most appropriate probe region and restriction endonucleases to use were deduced. The probe was amplified from the strain 4298 genome using primers ntf031 and ntf034 to yield a 755 bp product derived from the repeat region of the *hsr* gene. Probe labelling and hybridisation was carried out as described in Section 2.7.4. Figure 3.4a illustrates the restriction map of the HSRL showing predicted sites of hybridisation to the probe for strain 4298. Table 3.4 lists the expected restriction pattern of the strain 4298 digests and compares it with the actual outcomes shown in Figure 3.4b.

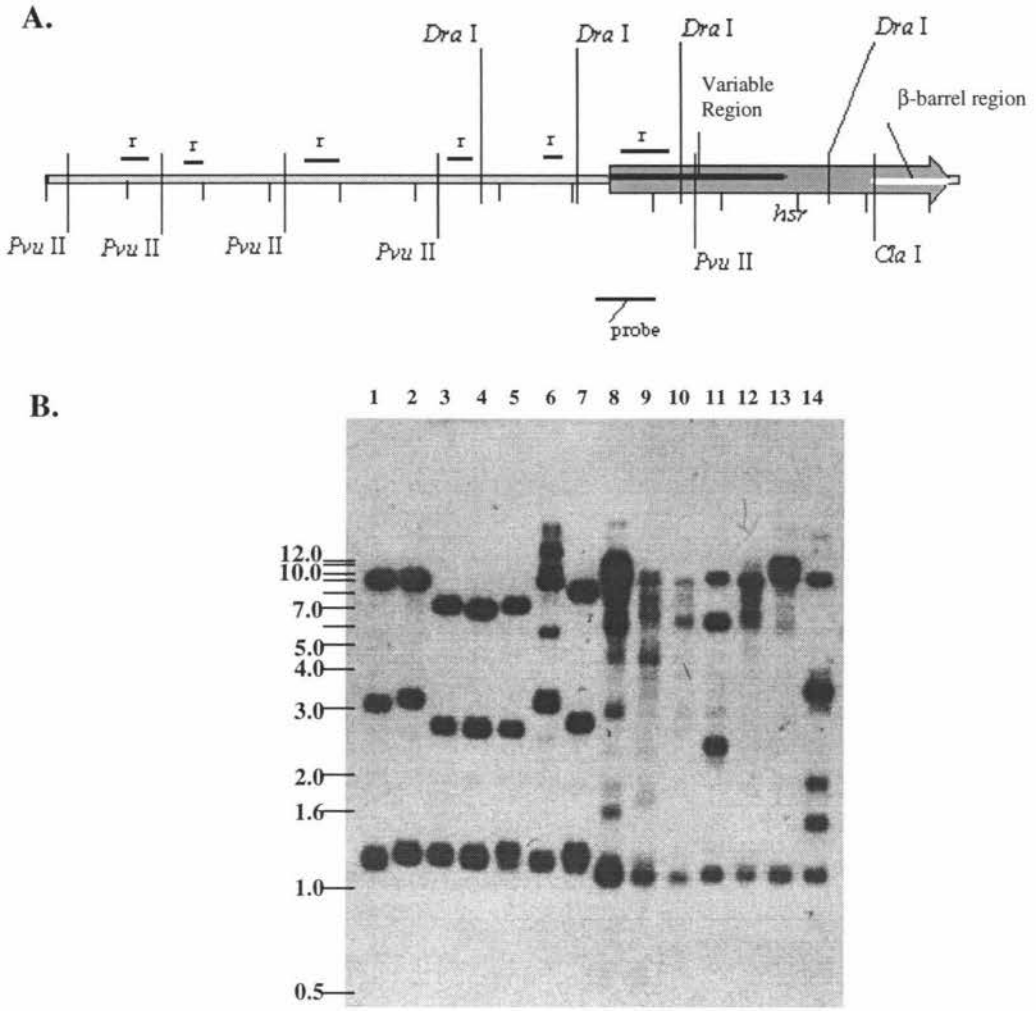


Figure 3.4 HSRL restriction patterns of *H. mustelae* isolates are different.

Panel A. Reference map of the 12144 bp *hsr* locus of strain 4298 for Southern hybridisation analysis. Restriction sites are labelled and indicated (refer to Table 3.5. for restriction analysis summary). The probe is shown as a labelled black line under the HSRL line. Black lines abbreviated with an 'r' are regions of homology where the probe was predicted to hybridise. The *hsr* gene is depicted as a light grey filled arrow, encoding the variable region (dark grey line) and the autotransporter region (white line). Lines below main HSRL line denote 1000 bp division markers.

Panel B. Southern blot and hybridisation with a 755 bp probe consisting of highly repetitive *hsr* DNA from strain 4298. Lanes: 1, F6 *Dra* I digest; 2, F7 *Dra* I digest; 3, F8 *Dra* I digest; 4, F11 *Dra* I digest; 5, F15 *Dra* I digest; 6, F21 *Dra* I digest; 7, 4298 *Dra* I digest; 8, F6 *Pvu* II/ *Cla* I digest; 9, F7 *Pvu* II/ *Cla* I digest; 10, F8 *Pvu* II/ *Cla* I digest; 11, F11 *Pvu* II/ *Cla* I digest; 12, F15 *Pvu* II/ *Cla* I digest; 13, F21 *Pvu* II/ *Cla* I digest; 14, 4298 *Pvu* II/ *Cla* I digest. The probe was prepared by PCR amplification of the *H. mustelae* reference strain 4298 DNA segment between primers ntf031 and ntf034 (HSRL coordinates 7330 – 8084).

Table 3.4 Restriction analysis and Southern blotting results for strain 4298

Restriction enzyme sets	Restriction sites in the 4298 HSRL (HSRL coordinates)	Restricted fragment sizes (bp)	Restriction fragments detected by Southern blotting. Approx. sizes (kb)
Dra I	5760	5760+*	1.3
	7040	1280*	1.4
	8411	1371*	2.8#
	10391	1980	8.5
		1752+	
Pvu II/ *Cla* I	254	254+	1.2
	1490	1236*	1.6
	3130	1640*	2.0
	5163	2033*	3.4
	8613	3450*	11.0#
	11011	2398	
		1028+	

* Expected to hybridise with single stranded probe DNA. + At least the indicated size with regards to the sequence information available. # Additional bands to those expected.

Obvious restriction length polymorphisms (RFLPs) could be detected in the banding patterns (Fig. 3.4b) from both restriction digest sets. The control digests resulted in patterns for strain 4298 and the probe, which were as expected (Table 3.4). The patterns for the *Dra* I digested samples were clearly visible and seemed to be digested completely, with the exception of F21. On the other hand, the *Pvu* II/ *Cla* I digested samples appeared undigested or partially digested, with the exception of the reference strain 4298 and isolate F11, which showed clear RFLPs (Fig 3.4). The strains could be easily clustered into 3 groups based on the *Dra* I banding pattern, which correlated with the source of isolation, i.e. *Helicobacter mustelae* isolated from colony reared ferrets (F6, F7 and F21), wild ferrets (F8, F11 and F15), and the control strain 4298. Within these groups there were subtle differences that could be detected on the blots, but the pattern was clearly much more conserved within a group. The size of the *hsr* loci of the other strains was estimated to be up to 18 kb, by calculating the sum of all reacting DNA fragments and including a non-hybridising fragment in the non-repeated half of the *hsr* gene.

Hybridisation detected an additional unpredicted band in the strain 4298 digested genomic DNA samples. Bands of approximately 2.8 kb and 11 kb were discovered in the *Dra* I and *Pvu* II/ *Cla* I digests respectively, and were not predicted if the HSRL-like sequences were confined to the 12144 bp locus. These results implied the existence of *hsr*-like sequences beyond the 12 kb HSRL and further sequence information was required to complete the total HSRL DNA sequence. Collectively, the results strongly

suggested that the repeat-containing regions of the HSRL varied in each of the seven strains. This arrangement was expected if the HSRL undergoes genetic rearrangements and justified the onset of a larger sequencing project to investigate the nature of the rearrangements in detail.

3.2.4 DNA sequence analysis of the variability in the repeat region of the HSRL of different *H. mustelae* strains.

3.2.4.1 PCR amplification and DNA sequencing of sections of the hsr gene from nine H. mustelae strains.

To investigate the variability of the *hsr* gene, three PCR fragments amplified from each *H. mustelae* strain were sequenced and aligned to the corresponding sequence of other strains. The strains included six New Zealand *Helicobacter mustelae* isolates plus three strains isolated from American ferrets – strain 4298, strain *Hm*180, and strain *Hm*181. The amplified DNA segments represented a portion of each of the three distinct regions of the *hsr* gene: the repeat region, the β-barrel region and the central region comprising the remainder of the *hsr* gene sequence. To amplify segments from the autotransporter domain and the central domain, oligonucleotides chosen by screening the DNA sequence of the HSRL of strain 4298 for appropriately placed priming sites flanking the area of the *hsr* gene to be amplified. The primer pairing details and PCR product sizes for 4298 are listed in Table 3.5. The DNA was amplified as described in Section 2.8 using appropriate primer pairs for defined reaction sets shown in Table 3.5 and displayed in Figure 3.5.

Table 3.5 Summary of PCR products amplified from three different regions of the HSRL of nine strains of *Helicobacter mustelae*.

Primers used to amplify DNA segments	Region represented by PCR product	Area amplified (HSRL coordinates, 4298)	Size of PCR fragment expected for strain 4298 (bp)	Sequencing primer used	Strains sequenced
ntf005/ntf034	Upstream region	958 – 1758	801	ntf034	F6
ntf005/ntf034	Repeat region	7575 – 8084	510	ntf034	All
ntf009/ntf026	Central region	9823 – 11015	1193	ntf009	All
ntf008/ntf037	β domain	11016 – 12101	1086	ntf037	All

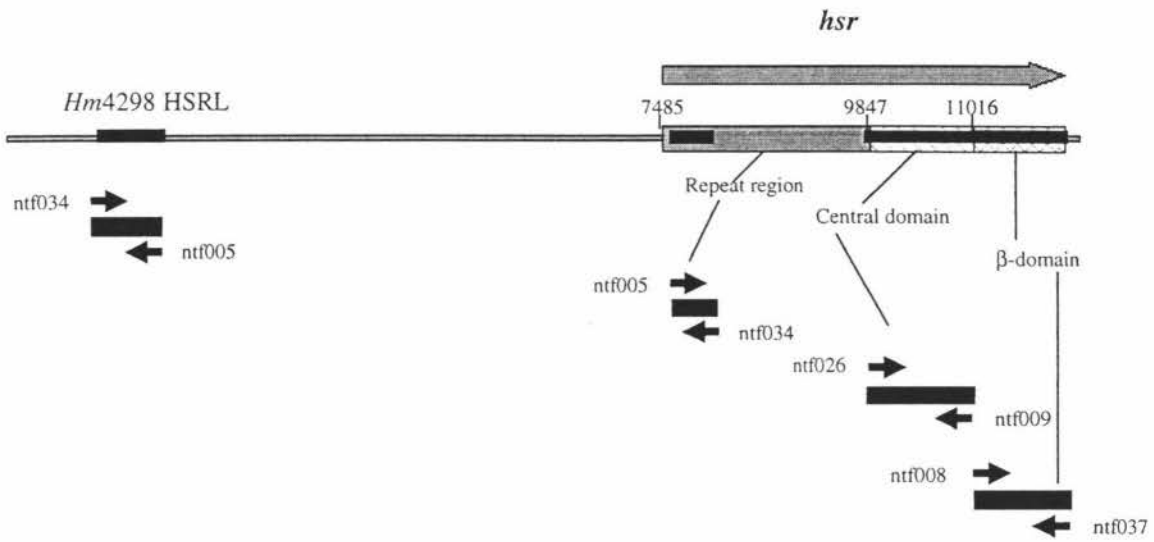


Figure 3.5 Comparative DNA sequence analysis of three regions of the *hsr* gene of nine *H. mustelae* isolates.

Fig. 3.5 shows the relative priming position of oligonucleotides with respect to the HSRL of strain 4298. Black boxes represent PCR products amplified with corresponding primers indicated by labelled arrows flanking the enlarged boxes. Refer to Table 3.5 for details regarding HSRL coordinates. HSRL is shown as a thin grey filled box and the *hsr* gene is illustrated as a grey arrow. The three regions of the *hsr* gene are labelled and separated by different fill patterns. HSRL coordinates are indicated at the beginning of each of the three regions of the *hsr* gene.

Oligonucleotides that primed within repetitive stretches of DNA were used to amplify a portion of the repeat region of all nine isolates. A single PCR reaction using primers ntf005 and ntf034 generated two products that were amplified from all nine strains, with sizes of approximately 500 bp and 800 bp. The PCR fragments were purified from an agarose gel following electrophoresis as outlined in Section 2.7.1 and concentrated to 20 ng/μl by Speedvac (Section 2.7.4). The DNA was submitted for DNA sequence analysis (Section 2.7.5), and the resulting sequence information obtained for each isolate was edited using EditView (ABI Prism Software). The sequences from each region of the gene were then aligned for sequence comparison using ClustalW DNA alignment programme (Thompson *et al.*, 1994).

From the DNA sequence of the strain 4298 *hsr* locus, the 510 bp fragment was located

within the *hsr* gene (Table 3.5). The identity of the corresponding fragment of the other *H. mustelae* isolates was examined by sequencing the 500 bp and 800 bp DNA fragments from a single strain (F6), which confirmed that the position is the same as the corresponding sized fragment in strain 4298 (i.e.the 500 bp fragment). Alignment of the 800 bp DNA sequence of F6 with the HSRL sequence of strain 4298 showed 95.4 % identity to HSRL co-ordinates 996 – 1624 upstream of the *hsr* gene (refer to Appendix 4e for DNA sequence alignment). The 500 bp DNA fragments from the remaining PCR reactions were then sequenced as described above.

3.2.4.2 Sequence alignments of regions of the hsr gene from nine H. mustelae strains.

Table 3.6 summarises the results of the DNA alignments of the sequences of PCR fragments of the *hsr* gene from all strains. For physical DNA alignments of all three DNA segments refer to Appendix 4. Both the central domain and the autotransporter domain showed few differences in DNA sequence across all nine strains with values of 98 - 99.7% and 99.3 – 99.8% identity to strain 4298 respectively. In contrast, the repeat region sequence showed considerable sequence variation resulting in a range of 51.8 – 62.1% identity with strain 4298. The repeat region comprised blocks of conserved DNA alternating with stretches of variable DNA. These conserved-variable-conserved motifs (CVCs) continued throughout the length of the sequenced product. The variation in DNA sequence was the main feature of the repeat region. Consequently, this area of the *hsr* gene was renamed the variable repeat region.

Clustering of percentage identity values, corresponding to the country of isolation of strains, was detected in the DNA alignments for the β-domain and the central domain with the values being higher for the American strains than the New Zealand strains when compared to strain 4298. However, there was no clear grouping of values observed in the repeat (variable) region DNA sequence alignment as was the case noted with the central and autotransporter domains.

Table 3.6 Summary of DNA sequence alignment of three different regions of the *hsr* gene of nine *Helicobacter mustelae* isolates. Title of each region appears in bold capitals. The percentage identity to strain 4298 is shown in bold.

REPEAT REGION

Strain name	Length of test strain sequence	Length of control sequence aligned	Length of alignment	Number of gaps in alignment	Number of Identities to 4298	% Identity to 4298
4298	470	470	470	137	470	**100.0**
F6	428	460	597	169	330	**55.3**
F7	432	466	603	171	362	**60.0**
F8	458	454	591	133	341	**57.7**
F11	448	451	588	140	345	**58.7**
F15	523	459	596	73	370	**62.1**
F21	445	463	600	155	311	**51.8**
Hm180	491	453	590	99	331	**56.1**
Hm181	435	454	591	156	320	**54.2**

CENTRAL DOMAIN

Strain name	Length of test strain sequence	Length of control sequence aligned	Length of alignment	Number of gaps in alignment	Number of Identities to 4298	% Identity to 4298
4298	670	670	670	1	670	**100.0**
F6	639	639	639	0	631	**98.7**
F7	643	643	643	0	637	**99.1**
F8	377	377	377	0	371	**98.4**
F11	418	418	418	0	412	**98.6**
F15	559	559	559	1	548	**98.0**
F21	670	670	670	1	658	**98.2**
Hm180	378	378	378	0	377	**99.7**
Hm181	309	309	309	0	308	**99.7**

β-DOMAIN

Strain name	Length of test strain sequence	Length of control sequence aligned	Length of alignment	Number of gaps in alignment	Number of Identities to 4298	% Identity to 4298
4298	608	608	608	0	608	**100.0**
F6	608	608	608	0	605	**99.5**
F7	608	608	608	0	605	**99.5**
F8	608	608	608	0	606	**99.7**
F11	608	608	608	0	606	**99.7**
F15	608	608	608	0	604	**99.3**
F21	608	608	608	0	605	**99.5**
Hm180	608	608	608	0	607	**99.8**
Hm181	608	608	608	0	607	**99.8**

3.2.4.3. Search for conserved-variable-conserved motifs (CVCs) in the hsr-like sequence flanking the hsr gene of strain 4298.

The possibility that the DNA sequences flanking the *hsr* gene might act as a potential source of sequence variation in the first half of the *hsr* gene was investigated. The *hsr*-like sequence flanking the *hsr* gene and the DNA sequence generated from the 800 bp PCR product of F6 (Section 3.2.4.1) were examined for conserved-variable-conserved (CVC) motifs using computer programmes (Geneworks 2.5, Intelligenetics) as search engines. The completed 14919 bp sequence of the *hsr* locus (refer to Section 3.3) was available for this search.

Figure 3.6 shows the CVC motifs aligned beneath a multiple alignment of the translated DNA sequences of the nine *H. mustelae* strains. Alignment of the translated variable region sequenced showed the organisation of conserved and variable residues in a simpler display format (Fig. 3.6 and Appendix 4d). Fifteen CVC motifs were found in the translated *hsr* flanking sequence of strain 4298 and one in the sequenced 800 bp of strain F6 sequence (corresponding to HSRL 1001 – 1171). These sequences were all directly related to the part of the variable region within the *hsr* gene. Entire CVC motifs present in the translated flanking sequences of strain 4298 were also found in the amplified regions of some of the ferret isolates, but none were repeated exactly within the *hsr* gene of strain 4298. Likewise the CVC sequence found in the translated F6 sequence was not precisely the same as the corresponding CVC sequence in the variable region of strain F6, although the variant portions of the CVC sequences were found precisely duplicated in both strains F6 and 4298.

The repetitive nature of the individual conserved and variable blocks presented in Figure 3.6 within the HSRL were also investigated (Appendix 3). The results implied that the conserved sequences were repeated many times while variable sequences were generally not repeated within the HSRL of strain 4298. Table 3.7 shows that all or parts of the conserved DNA sequences were repeated 3 - 9 times within the HSRL of strain 4298 whereas the variable sequences were not repeated with the exception of block V5 (Fig. 3.6), which was repeated three times.

Figure 3.6 Conserved-variable-conserved motifs in the *hsr* locus. Multiple alignment of part of the variable region of the Hsr proteins of nine *H. mustelae* strains, corresponding to residue 37 through 183 of the Hsr protein of strain 4298. Conserved sequences are boxed. Underlining indicates variant sequences present in the amplified regions of the ferret isolates, which are also found in the translated flanking sequences of strain 4298 or the F6 upstream sequence. Labelled grey boxes denote blocks of conserved (C1 - 6) or variable (V1 – 5) sequence. Arrowhead denotes signal peptidase cleavage site of Hsr of strain 4298. Modified from Forester *et al.*, 2001.

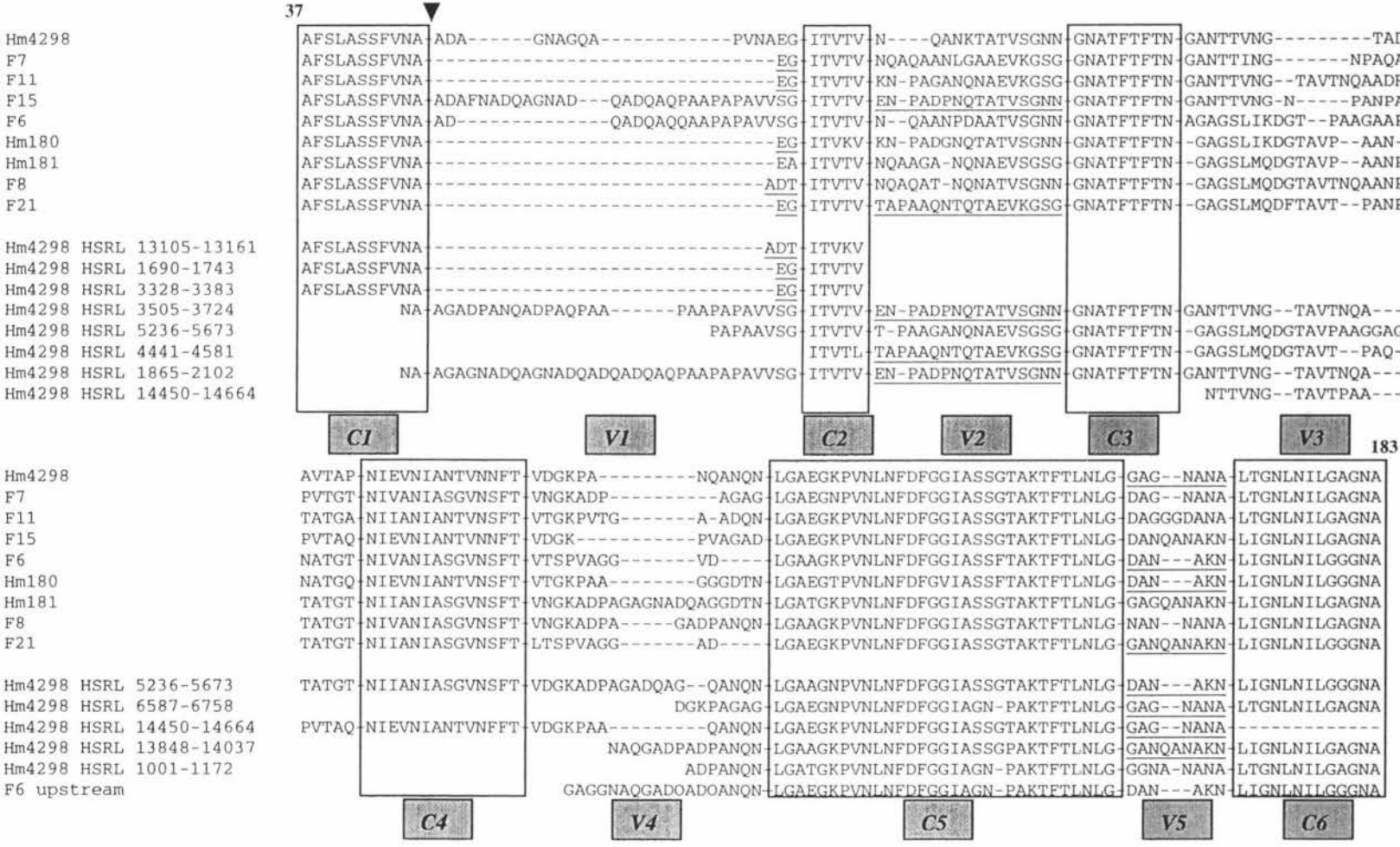


Table 3.7 **Repetitive nature of the conserved and variable sequence blocks from the repeat region of the *hsr* gene of strain 4298.**

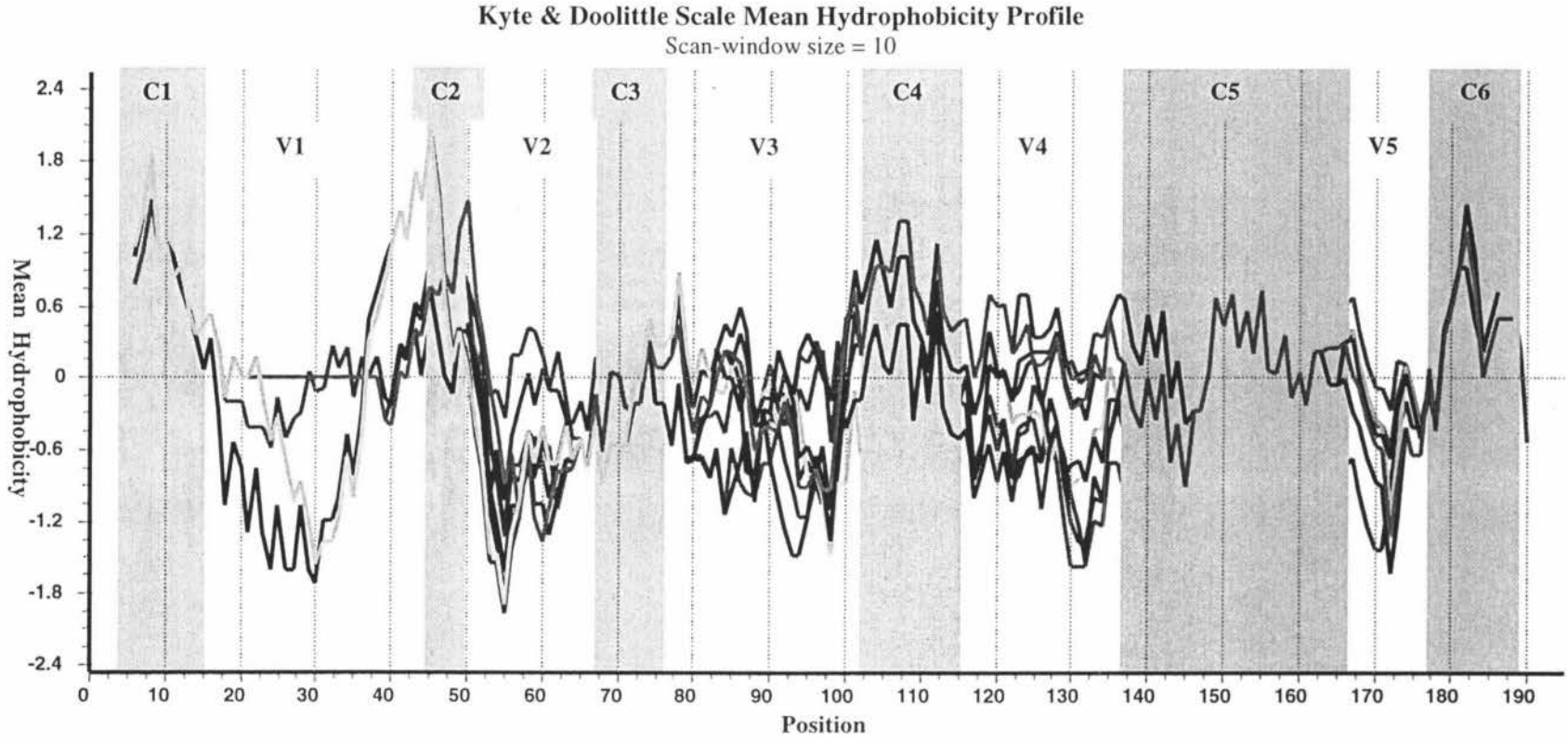
Block name	HSRL Co-ordinates	Length of sequence block	Relevant HSRL Co-ordinates repeated within the sequence block	Corresponding repeat No.s (Appendix III)	Length of repeat sequence	No. times repeated	Total No. times repeated[#]
C1	7579 - 7623	45	7579 – 7623	6	45	3	**4**
			7579 – 7623	8	45	2	
V1	7624 – 7664	40					**0**
C2	7665 – 7686	22	7665 – 7684	59	20	2	**7**
			7665 – 7683	109	19	3	
			7665 – 7682	119	18	5	
			7667 – 7686	25	20	2	
V2	7687 – 7698	12					**0**
C3	7699 – 7747	49	7699 – 7747	16	49	3	**6**
			7713 – 7728	148	16	4	
			7721 – 7747	54	27	5	
V3	7748 – 7778	31					**0**
C4	7779 – 7844	66	7799 – 7844	25	46	2	**3**
			7818 – 7833	140	16	2	
V4	7845 – 7877	33					**0**
C5	7878 - 7970	93	7878 – 7970	7	93	2	**9**
			7889 – 7930	35	42	4	
			7896 - 7968	19	74	3	
			7896 - 7930	50	35	3	
			7910 - 7928	108	19	7	
			7942 – 7970	52	29	3	
			7947 – 7970	5, 66	24	2	
V5	7971 - 7996	26	7971 – 7996	5	26	2	**3**
			7979 - 7996	22	18	3	
C6	7997 – 8046	50	7997 – 8046	5	50	2	**4**
			7997 – 8046	22	50	3	
			8007 – 8046	38	40	3	

[#]Refers to the number of times repeated less overlapping sequences common between repeat group numbers.

3.2.4.4. Properties of the peptide encoded by part of the variable repeat region of the hsr gene of nine H. mustelae strains.

The hydrophobicity profiles of conserved and variable blocks of protein sequence translated from the DNA sequence of the sequenced variable region segments were created using the Kyte and Doolittle scale mean hydrophobicity algorythm (Kyte and Doolittle, 1982). Figure 3.7 shows an overlay of the nine *H. mustelae* protein profiles. The conserved blocks were mostly hydrophobic, while the variable blocks appeared generally hydrophilic in nature.

Figure 3.7 Kyte & Doolittle scale mean hydrophobicity profiles of part of the variable protein sequences of nine *H. mustelae* isolates. Figure shows an overlay of the mean hydrophobicity profiles of sequenced part of variable region of nine *H. mustelae* strains. The window size is 10 residues. Conserved blocks of protein sequence are shaded in grey and labelled, C1 – C6. Variant protein groups labelled, V1 – V5.



3.3 Completion of the DNA sequence of the HSRL

The detection of an unexpected band in each of the 4298 digests following Southern blot analysis implied there was more *hsr*-like DNA in the 4298 genome (Section 3.2.3). The DNA sequence in the region directly downstream of the *hsr* gene was determined to find *hsr*-related sequences. This area had not yet been investigated, and appeared to be the most logical site to begin looking for the remaining *hsr*-like DNA sequence.

3.3.1 The lambda clone λE2

A recombinant lambda phage, λE2, expressing the intact Hsr protein from *H. mustelae* strain 4298 was previously constructed by Dr P.W. O'Toole and DNA from λE2 was available at the onset of this study (O'Toole *et al.*, 1994; Section 1.3.3). The λE2 clone was generated as follows. An approximately 23 kb genomic library fragment, generated by partial digestion of the 4298 genomic DNA with the restriction enzyme *Sau* 3A, was ligated between the left and right *Bam* HI digested arms of the lambda replacement vector, EMBL3. The resulting lambda clone was selected for its ability to produce a protein, which reacted with polyclonal rabbit antiserum raised against the purified Hsr protein. The clone DNA was then subjected to restriction analysis, subcloning and then DNA sequencing to determine the *hsr* gene sequence (O'Toole *et al.*, 1994). The sequence was submitted to GenBank and assigned the accession number, L15629. The initial restriction mapping of the λE2 DNA indicated that more genomic DNA, downstream of the *hsr* gene, was available for sequence determination (P.W. O'Toole, unpublished work) in this study.

3.3.2. Restriction endonuclease mapping and subcloning of λE2 for DNA sequence determination 3' of the *hsr* gene of *H. mustelae* strain 4298

Further restriction analysis was performed during this study to find appropriately sized and positioned DNA fragments to undergo DNA sequence analysis, such that background signals due to multiple priming sites were minimised. Approximately 200 ng of λE2 DNA was digested with *Bam* HI, *Eco* RI, *Hind III*, *Pvu* I or combinations of

these enzymes (not shown). A restriction map for λE2 was determined and is presented in Figure 3.8a (Appendix 1.1). The *Eco* RI digested λE2 DNA that gave a relatively simple spread of six DNA fragments (26.5, 17.0, 2.2, 1.9, 1.7 and 1.7 kb). Using previous sequence information as a reference, both the 1.9 kb and one of the 1.7 kb DNA fragments appeared to originate from within the *hsr* gene. The two larger fragments 26.5 kb and 17.0 kb could have only arisen from the ends of the clone, containing the flanking EMBL3 lambda arms. After cross-referencing the restriction patterns obtained in the other DNA digests, the two remaining fragments (1.7 kb and 2.2 kb) were mapped to the 3' end of the *hsr* gene and were suitably sized for subcloning and DNA sequencing (Fig. 3.8a).

Approximately 2 μg of λE2 DNA was digested using *Eco* RI enzyme. Following cloning strategies described in Section 2.9, the DNA fragments ranging from 1.7 kb - 2.2 kb *Eco* RI fragments were collected and purified from a 1% TAE agarose gel and ligated to *Eco* RI cut pUC19 vector in a 10 μl total ligation volume (Section 2.9.2). Half of the ligated reaction mixture was used to transform competent *E. coli* strain ER2206 (Section 2.9.3) and recombinants were selected for by blue/white selection. Easy preparations of plasmid DNA were made from twenty "white" recombinants (Section 2.6.1.1). Of those, 12 contained 1.7 kb and four contained either 1.9 or 2.1 kb *Eco* RI fragments. The recombinants were additionally digested with *Xba* I to both confirm the identity of the 2.1 kb insert and to determine its orientation with respect to the plasmid clone (not shown). The 2.1 kb *Eco* RI λE2 fragment was labelled the λE2/E1 DNA fragment and the clone containing it was called pUC19-λE2/E1. Likewise, clones containing the 1.7 kb fragment located 3' of the *hsr* gene (i.e., λE2/E2 DNA fragment), were named pUC19-λE2/E2. These clones were differentiated from those containing the *hsr*-derived 1.7 kb fragment by *Bam* HI digestion, since the latter did not contain this restriction site. Purified plasmids pUC19-λE2/E1 and pUC19-λE2/E2 were isolated from one of each clone using the CONCERT™ rapid plasmid purification miniprep system (Life Technologies) and prepared for use as sequencing templates as outlined in Section 2.7.5.

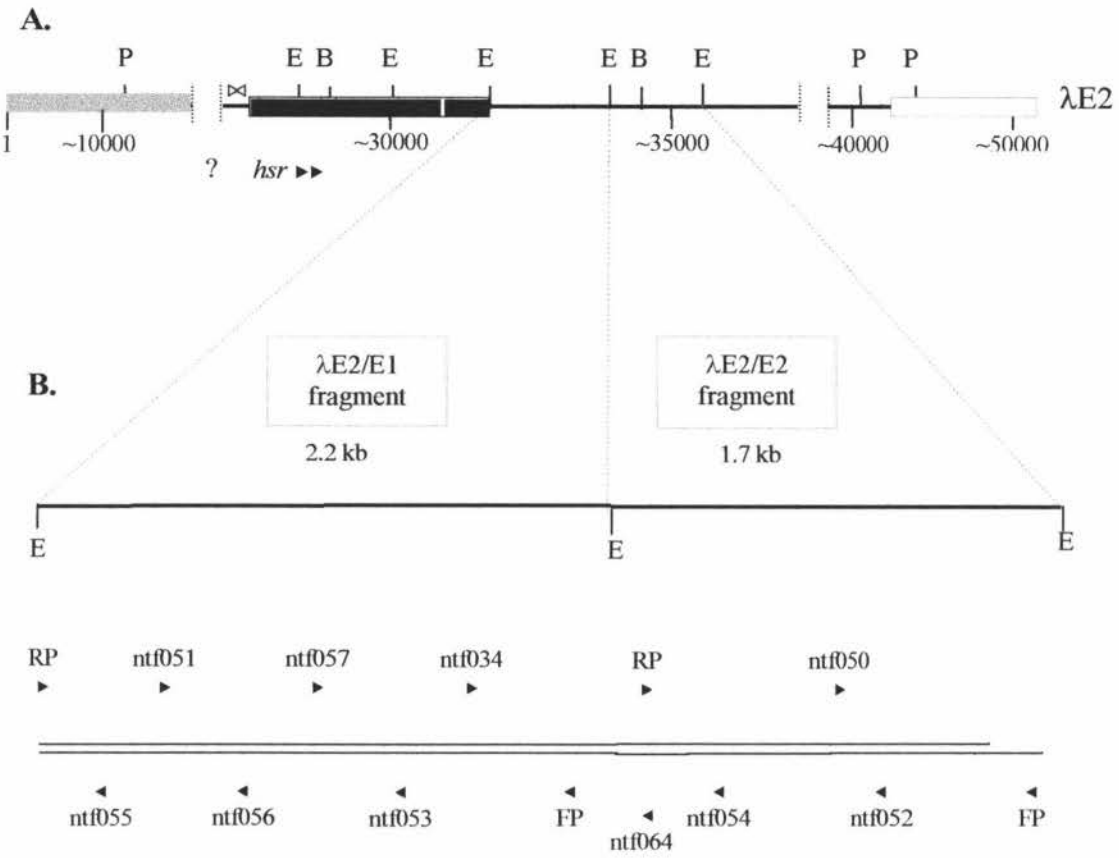


Figure 3.8 Restriction endonuclease and DNA sequence mapping of λE2.

A. The restriction map of λE2 is drawn schematically. Markings are shown for approximate co-ordinates along the length of λE2 (not to scale). The breaks in the λE2 line represent non-relevant sequence, which is omitted from this diagram. The grey rectangle to the left of λE2 depicts the 20 kb left arm, whilst the white rectangle depicts the 9 kb right arm of the lambda replacement vector EMBL3. The dark shaded box represents the *hsr* gene co-ordinates 135 – 4778 of Accession number L15629 (HSRL coordinates 7482-12039) and is labelled below with orientation arrows. The ⋈ denotes the scrambled region (Section 1.3.3). Unknown sequence upstream of the scrambled sequence region is shown by a ?. B, *Bam* HI; E, *Eco* RI; P, *Pvu* I.

B. Sequencing strategy for pUCλE2/E1 and pUCλE2/E2. Enlargement of DNA fragments λE2/E1 and λE2/E2 from panel A to show sequencing strategy. Areas sequenced on both strands are marked with a double line whilst those segments that were sequenced on one stand only are indicated by a single line. Primers used for sequence determination of the λE2/E1 and λE2/E2 fragments are indicated at their respective annealing positions. FP, pUC/M13 FP; RP, pUC/M13 RP.

3.3.3 DNA sequence analysis of pUC19-λE2/E1

The 2207 bp DNA sequence of the entire λE2/E1 DNA fragment was determined by primer walking using automated sequencing (Section 2.7.5), following the confirmation of the presence of *hsr*-like DNA beyond the *hsr* gene (using M13/pUC RP, Fig 3.8b). The DNA sequences were examined for accuracy and edited using EditView software tools (Section 2.7.5). BLAST nucleotide similarity searches of reliable DNA sequence detected *hsr*-related sequence in all sequencing reactions. Figure 3.8b shows the strategy for DNA sequence determination of the pUC19-λE2/E1 plasmid insert. The complete 2207bp λE2/E1 insert was sequenced on both strands of DNA. The λE2/E1 contig was generated from the individual sequences using the algorithm "sequencing project" of the Geneworks 2.5 software (Intelligenetics). A DNA dot matrix of the λE2/E1 contig, plotted against the 12 kb HSRL of strain 4298, showed extensive stretches of identity with the *hsr* locus of strain 4298 (not shown), and revealed that *hsr*-related sequence spanned the entire length of the λE2/E1 fragment. These results indicated that *hsr*-like sequence might extend further downstream of the λE2/E1 sequence. Consequently, DNA sequencing of the pUCλE2/E2 clone, comprising the downstream 1.7 kb *Eco* RI fragment, was required to complete the DNA sequence of the *hsr* locus of strain 4298.

3.3.4 DNA Sequence analysis of pUC19-λE2/E2

3.3.4.1 Sequencing of the hsr-like sequence in the λE2/E2 DNA fragment

The 1693 bp sequence of the λE2/E2 DNA fragment was generated by primer walking, and automated sequencing, as outlined for the λE2/E1 insert fragment in Section 3.3.3. Preliminary sequencing reactions employing M13/pUC forward and reverse primers on the pUC19-λE2/E2 template, followed by BLAST identity searches, showed two things. First, the verification of the presence of further hsr-like sequence, and second, that the boundary of the *hsr*-like repeated sequences occurred somewhere within this fragment,

since BLAST identity searches did not detect *hsr*-like DNA in the distal end of the λE2/E2 *Eco* RI fragment. Consequently, the DNA sequence of λE2/E2 fragment was not completed double stranded. Figure 3.8b shows the relative positions of the primers used to generate the DNA sequence of the λE2/E2 fragment. A 1693 bp contiguous sequence was assembled (Geneworks 2.5, Intelligenetics) containing 8 ambiguous nucleotides.

3.3.4.2 Other non-hsr related features of the λE2/E2 DNA fragment

A nucleotide BLAST search of the λE2/E2 DNA sequence beyond *hsr*-like DNA showed no related sequences. However, translated BLAST searches found significant matches to the glutamate racemase of *Helicobacter pylori* strains CCUG 17874 (AAC44708), 26695 (AAD07615), and J99 (AAD06074). The relevant open reading frame from the λE2/E2 DNA sequence was aligned against all three proteins identified in the database search. The resulting protein alignment, produced using the BioEdit sequence alignment editor (Hall, 1999), is shown in Figure 3.9.

Figure 3.9 Multiple alignment of the λE2/E2 ORF against *H. pylori* Glr proteins.
Identities shaded in dark grey and similarities in light grey.

```
10        20        30        40        50        60        70
Glr CCUG17874   MKIGVFDSGVGGFSVLKSLLKAQLFDEIIYYGDSARVPYGTKDPTTIKQFGLEALDFFKPHQIELLIVAC
Glr Hp26695     MKIGVFDSGVGGFSVLKSLLKAQLFDEIIYYGDSARVPYGTKDPTTIKQFGLEALDFFKPHQIKLLIVAC
Glr J99         MKIGVFDSGVGGFSVLKSLLKARLFDEIIYYGDSARVPYGTKDPTTIKQFGLEALDFFKPHEIELLIVAC
λE2/E2 ORF      LKLGIFDSGAGGLSVLEHVLRAEIFDSIIYYGDTARLPYGTKHPDSIICFCLEALEFLLVQNVDMIIVAC

                      80        90        100       110       120       130       140
Glr CCUG17874   NTASALALEEMQKHS-KIPIVGVIEPSILAIKRQVKDKNAPILVLGTKATIQSNAYDNALKQQGYLNVSH
Glr Hp26695     NTASALALEEMQKHS-KIPVVGVIEPSILAIKRQVKDKNAPILVLGTKATIQSNAYDNALKQQGYLNVSH
Glr J99         NTASALALEEMQKYS-KIPIVGVIEPSILAIKRQVEDKNAPILVLGTKATIQSNAYDNALKQQGYLNISH
λE2/E2 ORF      NTASAHALDAMHKAAPKIPIIGVIEPGILAIKNRLKNLDAKILVLGTKATIQSAQYQKHLQKLGYNNITA

                150       160       170       180       190       200       210
Glr CCUG17874   LATSLFVPLIEESILEGELLETCMRYYFTPLKILPEVIILGCTHFPLIAQKIEGYFMEHFALSTPPLLIH
Glr Hp26695     LATSLFVPLIEESILEGELLETCMRYYFTPLEILPEVVILGCTHFPLIAQKIEGYFMEHFALSTPPLLIH
Glr J99         LATSLFVPLIEESILEGELLETCMHYYFTPLEILPEVIILGCTHFPLIAQKIEGYFMGHFALPTPPLLIH
λE2/E2 ORF      IPTSLFVSLVEEGIFEGPLVEECLRYYFGGIDFVPDAIILGCTHFPLLQKPIAAYF------QNKSLLIH

                220       230       240       250       260                   309
Glr CCUG17874   SGDAIVGYLQQK---YALKKNAHAFPKVEFHASGDVIWLEKQAKEWLKL
Glr Hp26695     SGDAIVEYLQQN---YALKKNACAFPKVEFHASGDVVWLEKQAKEWLKL
Glr J99         SGDAIVEYLQQK---YALKNNACTFPKVEFHASGDVIWLERQAKEWLKL
λE2/E2 ORF      AGEAIVQYITQNSHLLFESKKSRSLLSTKAGIVGDLVDGAHKLDHKTCNKAAQ           AKNFTIL
```

The alignment showed 54.1% identity and 78.8% similarity of the λE2/E2 ORF to the three *H. pylori* proteins over a length of 220 residues (Fig. 3.9). The open reading frame was apparently larger than for the *H. pylori* proteins at 309 amino acids, of which only 220 amino acids could be aligned. Ambiguities in the single stranded DNA sequence derived from the distal end of the λE2/E2 may have contributed to the added length of this ORF.

3.3.5 Sequence arrangement with respect to the *hsr* locus of strain 4298 and completion of the 14919 bp *hsr* locus DNA sequence

According to the restriction map of λE2, the λE2/E1 and λE2/E2 fragments were located directly adjacent at the end of the *hsr* gene (Fig. 3.8a). However, an additional PCR abridging λE2/E1 and λE2/E2 was sequenced to discount the possibility of missing nucleotides from between the λE2/E1 and λE2/E2 sequence. Primers ntf034 and ntf054 were used to amplify an approximately 1.5 kb band from strain 4298 genomic DNA, which was purified as described in section 2.9.1. Sequencing primer ntf064 was employed to generate the DNA sequence, which was then aligned to the λE2/E1 and λE2/E2 contigs. Multiple alignment of the DNA segments (not shown) confirmed that λE2/E1 and λE2/E2 are located directly adjacent to each other in the 4298 chromosome.

PCR reactions were also performed to confirm the position of the λE2/E1-λE2/E2 contig with respect to the *hsr* gene of strain 4298. The λE2/E1 and λE2/E2 fragments together comprised 3901 bp of DNA sequence, 2880 bp of which belonged to the HSRL. The sequence of the 2880 bp fragment completed the HSRL, a total of 14919 bp in length. A list of the primers used and the expected sizes of the PCR products are shown in Table 3.8 and the resulting PCR products are shown in Figure 3.10. The resulting PCR product sizes agree with those expected from position of the λE2/E1-λE2/E2 contig at the 3' flank of the *hsr* gene. The published sequence of the *hsr* gene (accession L15629) also overlapped with the λE2/E1 sequence as expected. The complete 14919 bp DNA sequence of the strain 4298 *hsr* locus was submitted to Genbank and was assigned to the accession number AF254134.

Table 3.8 PCR confirmation of the arrangement of λE2 DNA with respect to the *hsr* locus (HSRL).

Area spanned by PCR reaction	Primer pair used	HSRL Co-ordinates	Expected size of PCR product (kb)	Calculated size from gel (kb)
hsr control	ntf011/ ntf037	11016 – 12101	1.086	1.1
hsr to λE2/E1	ntf011/ ntf055	11016 – 12747	1.732	1.7
λE2/E1 to λE2/E2	ntf034/ ntf052	13805 – 14919 (+ 1422 bp)	2.535	2.5
λE2/E1 to λE2/E2	ntf057/ ntf052	13265 - 14919 (+ 1422 bp)	3.075	3.0

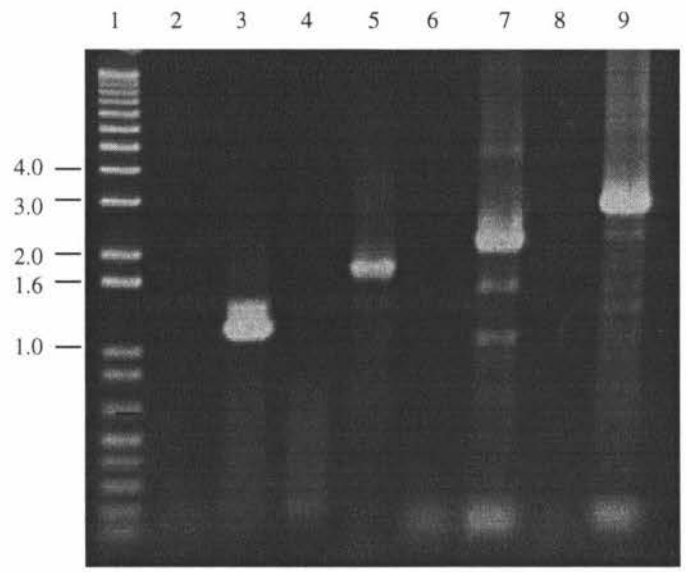


Figure 3.10 PCR confirmation of the assembly of the DNA sequence of the HSRL of strain 4298.

PCR analysis to confirm the placement of the sequenced λE2/E1 and λE2/E2 DNA fragments to the 3' flank of the *hsr* gene of strain 4298. Lanes 1, 1 kb plus size marker; 2, negative (water) control ntf011/ 37; 3, positive (DNA) 4298 control ntf011/ ntf037; 4, negative (water) control ntf011/ ntf055; 5, 4298 ntf011/ ntf055; 6, negative (water) control ntf034/ ntf054; 7, 4298 ntf034/ ntf054; 8, negative (water) control ntf057/ ntf052; 9, 4298 ntf057/ ntf052. Molecular sizes are indicated and are given in kb.

3.4. DNA sequence analysis of the 14919 bp *hsr* locus (HSRL)

3.4.1 Open reading frame analysis

101 potential open reading frames were identified in the *hsr* locus using a MacVector open reading frame analysis programme and applying the following criteria: 20 amino acids minimum, with ATG translational start site. Subsequently, the DNA sequence around the open reading frame was screened to find putative ribosome binding sites (RBS). Two mismatches in the standard "AGGAGG" RBS sequence were tolerated, and a maximum distance of 8 nucleotides between the translation start and RBS was allowed. Twenty-one conforming open reading frames were found and labelled in order from Orf 1 to Orf 21 according to size (Table 3.9). Each open reading frame was subjected to BLAST protein searches (BLASTp) to find related proteins.

Table 3.9 Open reading frames of the *hsr* locus.

ORF #	HSRL Co-ordinates start	HSRL Co-ordinates end	Length (bp)	Length (AA)	HSR Reading frame	Length between ATG and RBS (bp)	RBS sequence
1	7482	12038	4557	1517	3	7	AGGAGA
2	6445	5858	588	196	-2	6	AAAAGG
3	2227	1838	390	130	-3	4	AGGAAA
4	3849	3478	372	124	-1	4	AGGAAA
5	1609	1298	312	104	-3	5	AGGCGG
6	3249	2947	303	101	-1	8	AGCAGG
7	7207	7028	180	60	-3	5	AGATGG
8	346	513	168	56	1	3	AGGAGT
9	3271	3107	165	55	-3	4	AGGAAA
10	5242	5093	150	50	3	6	TGGTGG
11	5050	4910	141	47	-3	5	TGGTGG
12	10864	11001	138	46	1	5	AGCAGT
13	2944	2810	135	45	-3	8	TGGTGG
14	1338	1213	126	42	-1	4	AGGAAA
15	2084	2004	75	25	-1	8	AGCAGG
16	9139	9252	114	38	1	6	AGGCGG
17	13757	13647	111	37	-2	8	TGGTGG
18	12748	12855	108	36	1	2	AAGAGC
19	13726	13637	90	30	-3	5	AGGAAA
20	879	793	87	29	-1	5	AGGAAA
21	509	435	75	25	-2	7	AGGAGG

Orf 1 was the *hsr* gene and Orfs 3 - 21 showed either homology with the Hsr protein or no significant homology to any known protein. The only exception was, Orf 2 (196 amino acids). The BLASTp similarity search showed that it shared 41 and 40% identity

over 151 amino acids, with *Helicobacter pylori* LolA proteins HP085 (accession 025474, strain 26695) and jhp0722 (accession Q9ZL58, strain J99) respectively (Section 3.5.1.2 for details). *orf2* was translated from HSRL co-ordinates 6445 – 5858, located upstream of the *hsr* gene, interrupting the upstream repeat sequence region (USR) and in the opposite orientation to the *hsr* gene.

3.4.2 Repeat sequences of the HSRL

3.4.2.1 Analysis of dispersed DNA sequence repeats in the hsr locus

An analysis of the repetitive DNA occurrence in the HSRL of *H. mustelae* strain 4298 was performed using the DNA Strider (Dr Christian Marck, France) software package as described in Section 3.1 for the incomplete 12 kb HSRL for direct and inverted repeats. Search parameters were modified to a minimum length of 12 nucleotides with 100% identity. In total greater than 800 DNA sequence alignments were located, with 1046 stretches of repeated material comprising 343 perfectly repeated DNA sequences. These sequences were repeated from 2 to 11 times and were up to 741 nucleotides long. The repeats were categorised according to their location within the HSRL and summarised in Table 3.10. The detailed repeat co-ordinates are listed in Appendix 3. The relative distributions of the repeat sequences are illustrated in Figure 3.11.

An expansion of the existing repeat categories (Section 3.1) was done to take into account the extra DNA sequence contribution from the downstream flanking region. DSR was the collective term used to describe all repeats found in the downstream sequence regardless of repeat subtype. The downstream sequence repeats region (DSR) contained the following repeat subtypes. The downstream repeats (D) were found repeated only within the downstream of *hsr* (i.e., the DSR region), HD repeats were repeated at least once both downstream of and within the *hsr* gene, while HUD repeats were found in all three HSRL regions (i.e., USR, HSR and DSR, Fig. 3.11). The HD and HUD repeats are also a subset of the HSR group (Section 3.1). Subgrouping of the main repeat-containing regions of the HSRL facilitated the detection of repeats that were not in the *hsr* gene itself (i.e., U, D and UD repeats). These constituted a large

Table 3.10 Summary of the repetitive DNA sequences in the 14919 bp *hsr* locus. (HSRL).

Summary of the repeat type occurrence in the main regions of the *hsr* locus									
HSRL region	No. of repeats	No. of repeat stretches	Frequency of repeat type*						
			U	HU	H	UD	HUD	HD	D
USR	299	668	95	82	-	75	47	-	-
HSR	164	178	-	82	3	-	47	32	-
DSR	163	200	-	-	-	75	47	32	9
Total	**626**	**1046**	**95**	**164**	**3**	**150**	**141**	**64**	**9**
Actual No. repeats represented	**343**	**343**	**95**	**82**	**3**	**75**	**47**	**32**	**9**

Summary of repeat type occurrence in the *hsr* locus									
Number of times repeated	No. of repeats	Length range (bp)	Frequency of repeat type*:						
			U	HU	H	UD	HUD	HD	D
2	152	12 – 741	40	36	2	38	0	27	9
3	103	12 – 137	36	22	1	27	12	5	0
4	43	12 – 42	9	14	0	5	15	0	0
5	26	12 – 34	5	8	0	5	8	0	0
6	9	12 – 20	4	1	0	0	4	0	0
7	6	12 - 22	1	1	0	0	4	0	0
8	3	12 – 13	0	0	0	0	3	0	0
11	1	12	0	0	0	0	1	0	0
Total	**343**	**12 – 741**	**95**	**82**	**3**	**75**	**47**	**32**	**9**
Total repeat stretches	**1046**	**12 – 741**	**280**	**247**	**7**	**202**	**223**	**69**	**18**

*Repeat types: U, upstream repeat; HU, upstream and *hsr* repeat; UD, upstream and downstream repeat; HUD, upstream, downstream and *hsr* gene repeat; HD, downstream and *hsr* gene repeat; D, downstream repeat.

proportion (47.8%) of the total number of repeat material. In total 63.9% of the repeated sequence material was found in the USR, 19.1% in the DSR, and 17.0% in the HSR.

The complexity of the *hsr* repeats was demonstrated by DNA dot matrices generated using the DNA dot matrix programme of the Geneworks 2.5 software package (Fig. 3.12). The dot matrices show the alignment of the HSRL to itself, in the same and complementary orientations. A dot on the matrix alignment represents an 8-residue-identity unit. The matrices were compressed by a factor of eight for visualising purposes. The density of the 'dots' in the matrices positively correlated with the repeat concentration, and was exhibited for most of the USR and DSR regions. Likewise, the repeat (variable) region of the *hsr* gene showed dense clustering of dots upon alignment with the complement of the HSRL upstream and downstream flanking sequence. These results demonstrated that the *hsr*-like repeats were mostly inverted with respect to the *hsr* gene.

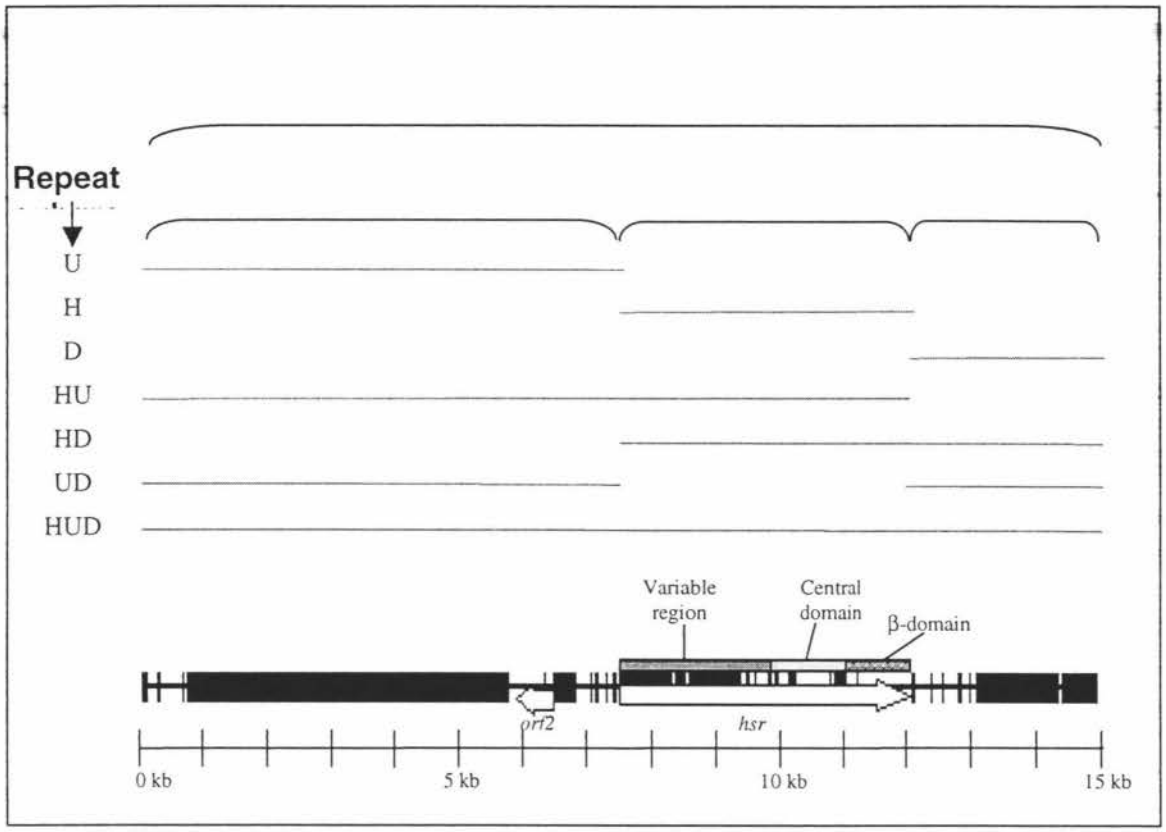


Figure 3.11 Repeat distribution in the *hsr* locus of *H. mustelae* strain 4298

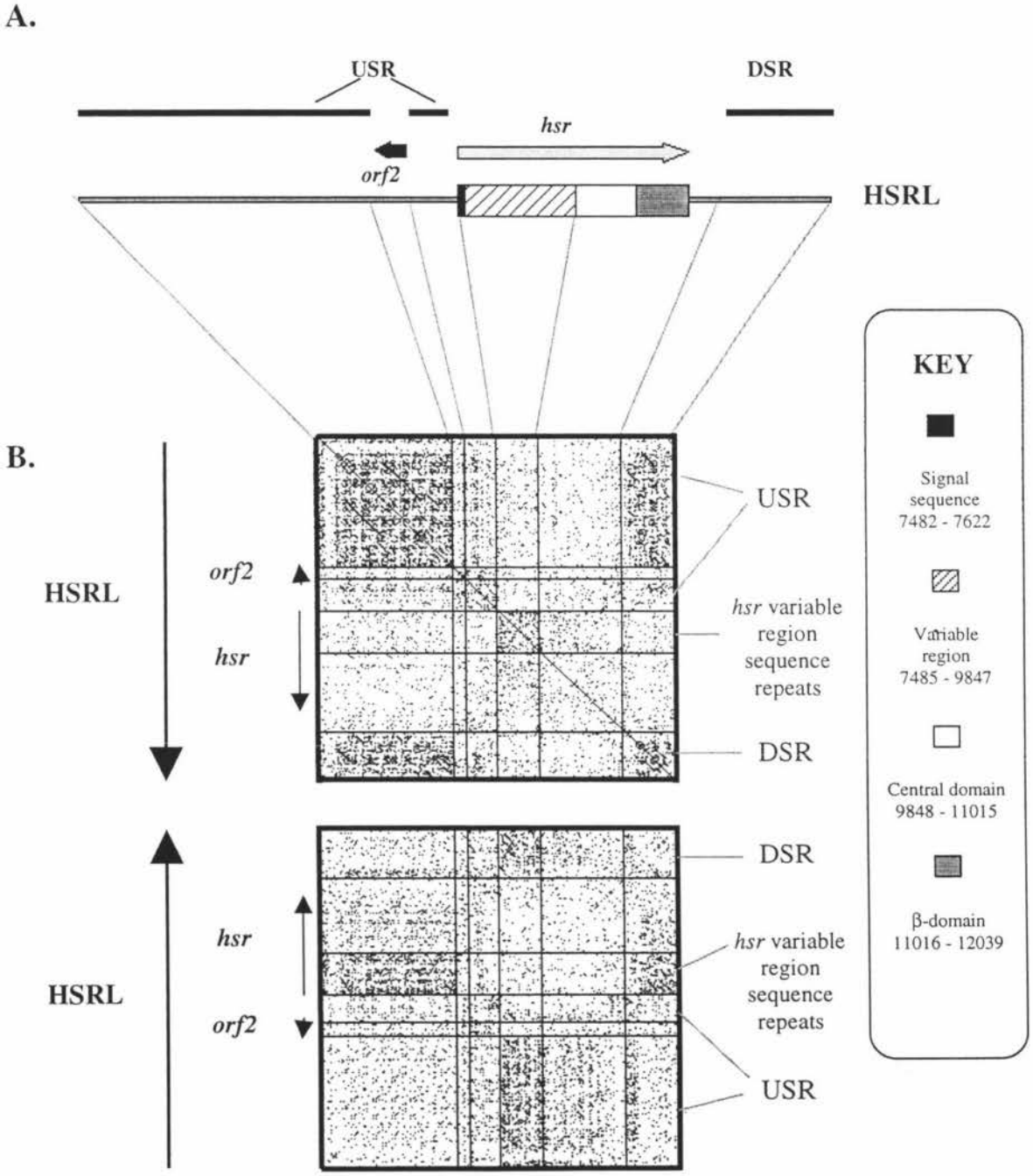
Top: The three major repeat-containing regions (USR, DSR, and DSR) of the HSRL, comprise repeats from different subgroups (indicated). The areas defining each repeat subgroup are represented as grey lines. **Bottom**: *H. mustelae* strain 4298 HSRL showing the repeat blocks (thick black lines or 'boxes') along the length of the HSRL. The *hsr* and *orf2* are shown as unfilled arrows. The three regions of the *hsr* gene are indicated. A scale line is included with each segment representing 1 kb.

A non-translated region of apparent low-dot density was displayed in the USR region corresponding to a break in the repeat sequences at HSRL coordinates 86-230 and 280-645 (Fig. 3.11, Fig. 3.12b and Appendix 3). Low dot density was also observed, in the same dot matrix, in the region containing *orf2*, the central and β-domains of the *hsr* gene, and in the 5' portion of the DSR.

The dot matrix of the HSRL vs HSRL complementary strand indicated the 5' DSR area contained direct repeats (Fig. 3.12b; Appendix 3). Eight direct *hsr*-like repeats (12-37 bp), with respect to the *hsr* gene, were found in the HSRL flanking the *hsr* gene (i.e. HU, HD or HUD repeats). All of these repeats were located at the 3' end of the *hsr*

Figure 3.12 Physical organisation and repeat sequences in the *hsr* locus (HSRL) of *H. mustelae* strain 4298.

(A). Line diagram showing the layout of the HSRL. A light grey arrow indicates the hsr gene and its orientation and *orf2* is shown as a solid black arrow. The domains of the *hsr* gene and relevant HSRL co-ordinates are shaded along the HSRL length (see figure key). USR, upstream sequence repeats; DSR, downstream sequence repeats. (B). Dot matrices showing the repetitive nature of the HSRL. Vertical lines connecting the HSRL organisation to the dot matrix layout point out relevant HSRL sections, which are labelled in panel A. The corresponding horizontal lines represent the same sections as in panel A, with most significant sections labelled. The orientation of the HSRL is shown as large arrows, whilst the individual gene orientations are indicated with a small arrow.



gene between HSRL co-ordinates 12041 – 12973 (933 bp).. Figure 3.13 shows the relative distribution of these directly repeat sequences with regard to the *hsr* gene.

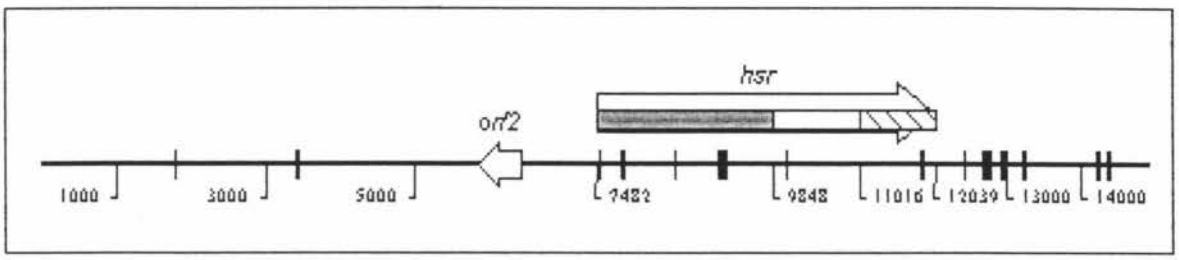


Figure 3.13 Distribution of the HSR direct repeats in the HSRL

The relative distribution of direct repeats containing at least one sequence within in the *hsr* gene. Genes are indicated and shown as unfilled arrows. The domains of the *hsr* gene are shown; variable region (grey), central region (unfilled), and the β-domain (diagonal lines). HSRL coordinates are given below the HSRL line.

Repeat sequences greater than 15 bp were not found in the central and β-domains. However, 14 smaller repeat stretches comprising 12 repeats were present in these regions, ten in the central domain and two in the β-domain. Ten from twelve of these repeats comprised sequences that were also repeated in the USR and DSR, while the remaining two H type repeats comprised three DNA stretches found outside the variable region (Fig 3.11).

3.4.2.2 Tandem repeats in the hsr locus

Repeat 273 (TAATGCTAATGC; Appendix 3) was repeated more often than any other repeat (11 times; Table 3.10), and appeared as a tandem repeat (TR) of a 6-bp sequence TAATGC that was repeated 35 times in the HSRL. The number of hexameric tandem units ranged from 2 – 4 units long. The self-aligned DNA dot matrix of the HSRL identified other apparent tandemly repeated nucleotides (6-18 bp) along the length of the HSRL and these TRs are listed in Table 3.11.

Three homopolymeric nucleotides greater than 8 nucleotides were also detected. All of these were located directly upstream of the *hsr* gene (Table 3.11).

Table 3.11 Tandem repeats in the *hsr* locus

TR#	TR sequence	TR unit length	No. times repeated in cluster	HSRL co-ordinates of TR	No. times TR units are repeated in the HSRL	Distribution (hsr gene: hsr flank)
1	GTGCAG	6	3	2711-28	40	6:34
			3	4333-50		
			2	4806-17		
2	TAATGC	6	2	1041-52	35	8:27
			3	2929-46		
			3	5035-52		
			2	6627-38		
			4	14145-68		
			2	14419-30		
			2	7990-79		
3	GATCCAGCA	9	2	2824-41	13	2:11
			2	4924-41		
4	TGATTTGCT	9	2	7856-73	11	3:8
5	ACCTGCTGC	9	2	9141-58	11	3:8
6	GATCAAGCA	9	3	2036-62	9	0:9
7	TAGGTGGTG	9	5	5695-740	7	0:7
8	TGGTGTAGG	9	5	5691-735	5	0:5
9	CAAATCCAG	9	3	2531-48	5	1:4
10	AGGAAACATTA	11	2	2654-85	5	1:4
			2	4286-307		
11	AGCTCCAGC	9	2	3659-76	4	0:4
12	GGTAATGCAGA TCAAGCA	18	2	2054-89	2	0:2
13	T	1	9	7396-404	1	0:1
			10	7435-44	1	0:1
14	C	1	10	7308-17	1	0:1

3.4.2.3 Imperfect inverted repeats and potential stem-loop structures in the hsr locus

A search for potential stem-loop structures (imperfect inverted repeats) was performed using DNA Strider hairpin search software (Dr Christian Marck, France). Search parameters were set to 12-nucleotide stem length, 2 mismatches were allowed and up to 16-nucleotide loop length. In addition, one 22-bp perfectly palindromic sequence was found manually during analysis of direct repeat region (Section 3.4.2.1). Thirteen PSLs were found in the HSRL with stem lengths ranging from 11 – 19 nucleotides (Table 3.12). The distribution of these stem loops, relative to the *orf2* and *hsr* gene, can be seen in Figure 3.14. PSLs were observed directly upstream of *orf2* (PSL 10) and *hsr* (PSL2 & 5).

Table 3.12 Potential stem loop (PSL) structures in the *hsr* locus

PSL#	HSRL coordinates	Length (bp)	Stem length	# matches in stem	Loop length	HSRL region: comment
1	240-88	49	19	17	11	USR: *
2	7373-406	34	14	2	6	USR-HSR junction: directly upstream of the *hsr* gene
3	1084-115	32	15	13	2	USR
4	6667-98	32	15	13	2	USR
5	7432-63	32	13	12	6	USR-HSR junction: directly upstream of the *hsr* gene
6	9093-123	31	14	12	3	HSR
7	540-69	30	12	10	6	USR: *
8	11415-443	29	12	10	5	DSR
9	514-41	28	14	12	0	USR: *
10	6456-82	27	13	11	1	USR: directly upstream of *orf2*
11	14196-222	27	13	12	1	DSR
12	9095-121	27	12	11	3	HSR
13	12571-92	22	11	11	0	DSR: found manually

* In the first low-dot density region of the USR region of strain 4298 HSRL sequence (Fig. 3.11b; Fig. 3.14).

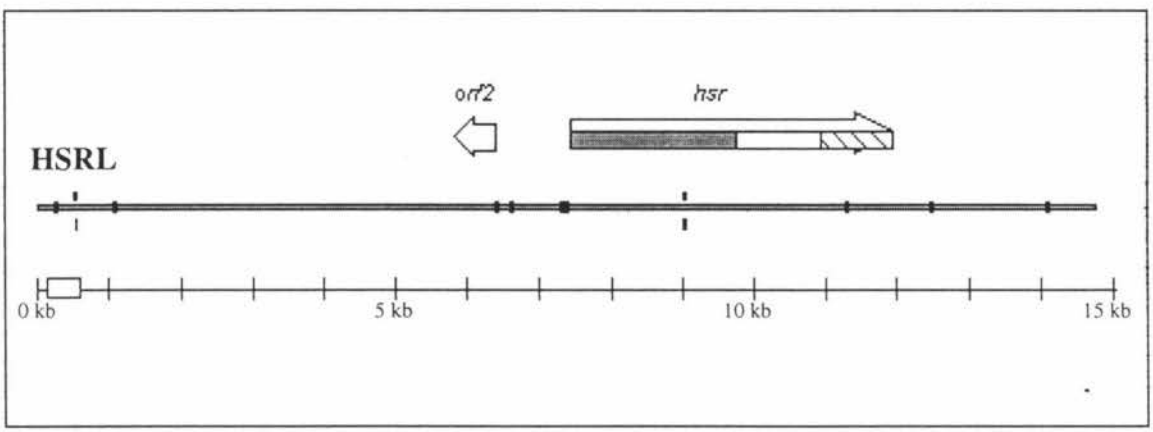


Figure 3.14 Distribution of potential stem-loop structures in the *hsr* locus

The relative positions of potential hairpin structures in the HSRL are shown (dense black blocks). Places where two hairpins are very close together or overlapping are shown with one PSL above and below the HSRL line. The *orf2* and *hsr* genes are shown as unfilled arrows. The three regions of the *hsr* gene are shown overlapping the *hsr* gene: variable region (grey), central region (unfilled), and the β-domain (diagonal lines). The low-dot density clearing (white box) is displayed (HSRL 86-230; 280-645) on the HSRL co-ordinate line.

3.4.2.4 Analysis of imperfect repeats in the hsr locus

An attempt was made to analyse repeats allowing mismatches using various computer analysis packages available in the lab and on the World Wide Web. None of the

programmes could accomplish this task successfully, as these repeats were apparently too numerous to cope with.

A conspicuously large approximately 1500 bp repeat was clearly visible in the dot matrix, and corresponded to HSRL coordinates 1311 – 2829 (1519 bp) and 2951 – 4451 (1501 bp) (Fig. 3.12b). This repeat was interrupted in both locations, by variable sequences, 56 bp and 38 bp in length respectively. Both variant stretches contained tandem repeats. TR6 and TR12 were housed in the first variant stretch, while TR11 was found in the second variant sequence. TR12 appeared to be a composite repeat containing a TR6 unit. There were also eight small single nucleotide changes within the large repeat.

3.5 Knockout mutagenesis of the *orf2* gene in strain 4298

3.5.1 Analysis of the *orf2* gene and gene products

3.5.1.1 The orf2 gene

The *orf2* gene was located within the *hsr* locus (HSRL 5858 – 6445), within the repeat array, upstream of the *hsr* gene. The *orf2* gene is 588 bp in length, in the opposite orientation to the *hsr* gene, and is preceded by a presumptive ribosome-binding site (AAAAGG, Section 3.4.1) located 7 nucleotides upstream of the translational start site. The sequence directly upstream of the *orf2* gene was examined visually for upstream elements. Hypothetical –10 and –35 promoter elements with two mismatches to the consensus *E. coli* promoter sequences (Harley and Reynolds, 1987) were located at HSRL 6463 – 6468 (AAAAAT), and HSRL 6452 – 6457 (TTCAGA) respectively (Fig. 3.15). The nucleotide composition of the putative *orf2* gene is GC-rich at 58.0%. A single 12-bp UD repeat unit was found within *orf2*, (i.e., repeat #314, HSRL coordinates 6317 – 6328). The next major repeat sequences flanking *orf2* occurred 118 bp (HSRL 5741) upstream and 28 bp (HSRL 6517) downstream of *orf2*. Terminator structures were not observed. Potential stem-loop structure, PSL5 (Section 3.4.2.3), is mapped to the *orf2* promoter region (HSRL 6456 – 6482) including the –10 sequence and two nucleotides of the putative ribosome binding site, but not the –35 promoter sequence (Fig. 3.15).

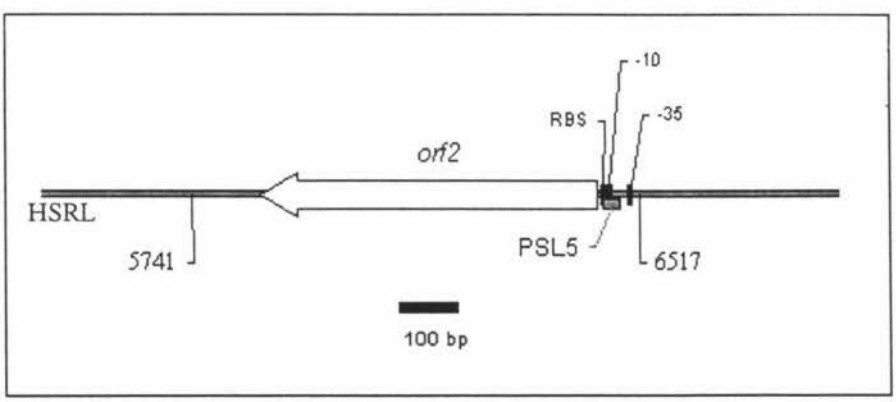


Figure 3.15 Local DNA features of the *orf2* gene

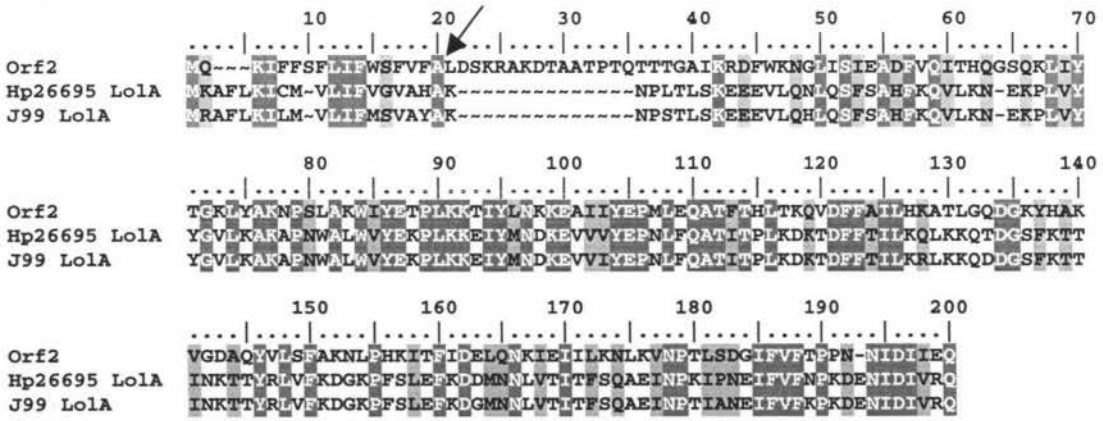
Figure shows the distribution of upstream elements (labelled) with respect to the *orf2* sequence. HSRL coordinates are shown below the main, grey *hsr* locus (HSRL) line.

3.5.1.2 The Orf2 protein

The putative Orf2 protein was 196 amino acids in length (22.4 kda) and displayed most similarity with the LolA proteins of *H. pylori* by BLAST protein database homology searches (Section 3.4.1). LolA is a periplasmic lipoprotein carrier protein (Matsuyama *et al.*, 1995). Forced alignment of the complete protein sequences showed the three proteins (Fig.. 3.16) share 34.1% identity and 53.5% similarity over 185 residues. The region of highest conservation showed as high as 58.6% identity over 58 residues. The signal sequence was predicted using SignalP (version 2.0) program (Nielson *et al.*, 1997) as the first 17 amino acids of the putative protein. The Orf2 protein was 12 amino acids larger than the LolA proteins of *H. pylori* strains J99 and 26695, with most of the extra sequence occurring at the N-terminus of the predicted mature Orf2 protein. The mature Orf2 protein contained a 14 amino acid N-terminal overhang, with respect to the *H. pylori* LolA proteins. This extra sequence contained mainly polar and charge residues.

Figure 3.16 Alignment of Orf2 translated product against *H. pylori* LolA proteins.

Protein names are shown to the left of the aligned sequences. Identities and similarities are highlighted in grey and light grey respectively. The arrow shows the putative signal peptidase cleavage sites of the aligned proteins.



The amino acid composition of the predicted mature (179 amino acids, 20.3 kDa) protein comprised 42.5% nonpolar/ hydrophobic amino acids, with a Kyte-Doolittle grand average of hydrophobicity (GRAVY) hydrophathy score (Kyte and Doolittle, 1982) of –0.293. The predicted mature protein also contained 31.8% polar uncharged, 10.1% acidic and 15.6% basic amino acids. The overall charge at pH 7 was +5.46, with an isoelectric point (pI) at pH 9.82 (Geneworks 2.5, Intelligenetics). The predicted

mature protein composition was dominated by four amino acids (Ile, Lys, Leu, and Thr) comprising 41.3% (41.8% MW) of the mature protein. Sulphur-containing residues comprised 0.6% of the mature protein, which was solely contributed by the f-methionine, and cysteine residues were absent.

3.5.2 Generation of the *orf2* knockout plasmid pHM205ΔORF2

A plasmid construct was made to disrupt the *orf2* gene of strain 4298 by inserting the *aphA3* gene into the *orf2* gene. Plasmid pHM205 (Appendix A1.15; Forester, previous work, unpublished) contained HSRL coordinates 3500 – 6462 (2961 bp *Eco*RI to *Sma*I fragment), which included the entire *orf2* gene, but not the upstream promoter elements. The *aphA3* gene was excised from plasmid pILL600 (Labigne-Roussel *et al.*, 1988) with *Sma* I and ligated to *Msc* I linearised pHM205, following cloning strategies outline in Section 2.9 to yield kanamycin resistant colonies at a frequency of approximately 4.5 x 10^5 cfu/μg of ligated DNA. Easyprep DNA from 12 kanamycin resistant transformants was *Hind* III digested for insert orientation determination. A representative colony housing plasmid with insert in either orientation was frozen for storage. pHM205ΔORF2.6 clone carried *aphA3* in the same orientation as *orf2*, as determined by restriction analysis, (using *Bam* HI, *Hind III,* or both), and was used for further transformations in *H. mustelae* strain 4298.

3.5.3 Transformation of *H. mustelae* strain 4298 with pHM205ΔORF2

H. mustelae strain 4298 was transformed with pHM205ΔORF2 as described in Section 2.9.3.3. The kanamycin concentration was determined by titration to 60 ug/ml final concentration in CBA as outlined in Section 2.2. Twenty kanamycin resistant, urease positive transformant colonies were analysed by PCR using primers KAN2 and ntf016 to generate a band of ~1.5 kb if the kanamycin cassette had been included into the 4298 genome. Nineteen colonies tested positive for the insertion (Fig. 3.17), of which, six were further analysed by protein electrophoresis on a 7.5% SDS-PAGE gel as described in Section 2.11.1. The resulting coomassie-stained SDS-PAGE gels demonstrated no detectable difference in the relative Hsr protein production between the strain 4298 transformants containing the *orf2* knockout and the wild type strain 4298 (Fig 3.8).

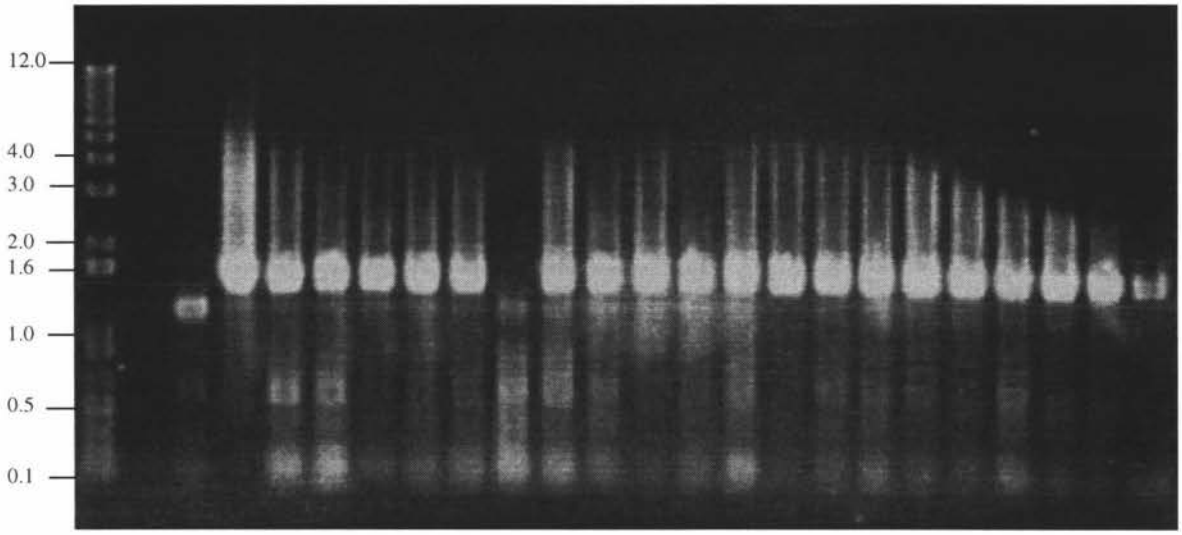


Figure 3.17 PCR confirmation of kanamycin resistant 4298ΔORF2 transformants.

PCR analysis of 20 kanamycin resistant 4298 transformants with primers ntf016 and KAN2, with an expected size of 1.5 kb for positive clones. Lanes: 1, 1kb plus DNA size marker (sizes given to the left of the gel in kilobases; 2, water only control; 3. 4298 control; 4 – 23 Kanamycin resistant clones 1 –20.

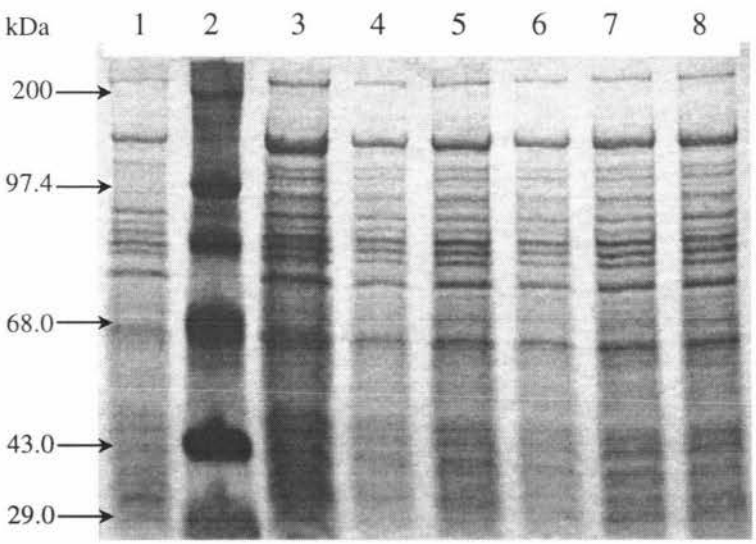


Figure 3.18 SDS-PAGE analysis of whole cell lysates of six kanamycin resistant 4298ΔORF2 transformants

Coomassie-stained polyacrylamide gel showing the relative expression of the Hsr protein in the reference (strain 4298) and kanamycin resistant *orf2* knockout transformants. Lanes: 1, 4298 cells; 2, Prestained protein molecular weight standard (Life Technologies); 3 – 8, kanamycin resistant, PCR positive clones #1 - 6.

4.0 DISCUSSION

4.1 *Helicobacter mustelae* is present in ferrets from at least two geographically distinct locations in New Zealand.

Successful culture of *Helicobacter* from three out of four breeding ferrets and three from seventeen wild New Zealand ferrets was achieved. The success of culture appeared to negatively correlate with the condition of the stomach. Furthermore, collection of ferret stomachs was (of necessity) carried out in the summer months, with warmer weather contributing to tissue decomposition, presumably rendering *H. mustelae* cells non-viable for primary culture.

H. mustelae has been detected in ferrets in all but one gastric microbiology investigation carried out so far. This study, reporting the absence of *H. mustelae* in a single population of ferrets from a pelt farm, in New Zealand, (Morris *et al.*, 1988) has been cited by others to indicate that this organism is not present in New Zealand (Fox and Lee, 1997). DNA analysis of the 16S rRNA gene was from a single sequencing reaction, as little as 16% of the total gene, but the region sequenced had been chosen to be sufficiently variable to distinguish between different strains of *Helicobacter*. The sequence identity values of the New Zealand isolates, to strain 4298, were from 98.3 – 100% (Table 3.3). Examination of the total cellular proteins of the ferret isolates on SDS-PAGE gels showed almost identical protein profiles, with all producing a prominent ~150 kDa protein, which was shown to react with the JA6 antiserum in a similar manner to that of the reference strain 4298. The results of the present study shows that *H. mustelae* is present in captive and wild ferret populations in two geographically distinct locations in the North Island of New Zealand.

4.2 Distribution of the *hsr*-related repeats

4.2.1 The *hsr* gene is flanked by multiple repeats of *hsr* sequence, in the 15 kb *hsr* locus of *H. mustelae* strain 4298

4.2.1.1 The distribution and complexity of repeat sequences in the 15

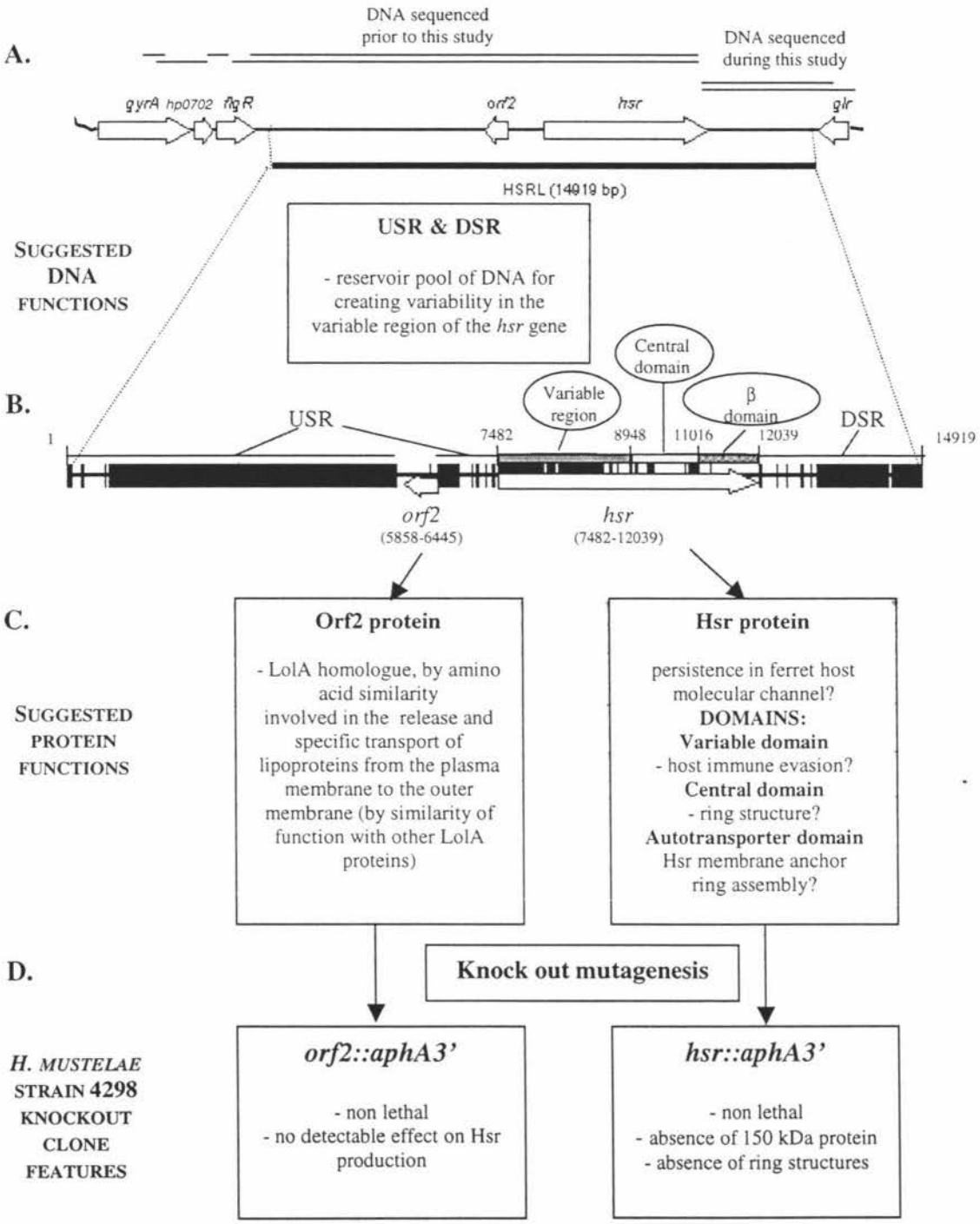
kb hsr locus of strain 4298

Southern hybridisation analysis with a probe comprising repetitive DNA demonstrated that all the *hsr*-like sequence was limited to a genomic area of no more than 16 kb in strain 4298, and up to 18 kb for the other *H. mustelae* strains examined. The first part of these results was confirmed by the completion of DNA sequence of the Hsr locus from strain 4298. The locus boundaries of strain 4298 established using BLAST nucleotide similarity searches were confirmed by dot matrix surveillance of all neighbouring sequenced DNA. The 14.9-kb *hsr* locus (Fig. 4.1) contains *hsr*-like sequences, and two open reading frames (*orf2* and *hsr*), which are flanked by a total of around 9.7 kb of multiple hsr-like sequences fragments. The genome region apparently dedicated to *hsr*-like sequence is relatively small, compared with the silent gene (*pilS*) repertoire encoding partial gonococcal pilin protein sequence (>35 kb) (Hamrick *et al.*, 2001, Section 1.5.2). The *hsr* flanking regions may represent an "economical" means for potential sequence variation using a relatively small part of the *H. mustelae* chromosome to generate large numbers of sequence variants (Section 4.4).

A total of 343 sequence stretches (12 – 741 bp) were perfectly repeated from 2 to 11 times. The size of the repeated sequences varied from 12 to 741 bp. Of these sequences 65% of the repeat material was found in the USR, 18.4% in the DSR, and 16.6% in the *hsr* gene (HSR), a relatively equal distribution of repeats considering the length of chromosome they cover (i.e. ~7 kb, 2.8 kb and 2.3 kb, respectively). The distribution and the complexity of the *hsr*-like repeats are well presented by dot matrices (Section 3.4.2). These shows a number of obviously large sequence repeats that were not detected by DNA Strider perfect repeated sequence search (e.g., a 1.5 kb imperfect duplication). However, compression of the sequence alignment compromised the resolution of the dot matrix, such that special features, for example, tandem repeats and homopolymeric tracts were not evident at low resolution. Despite an eight-fold compression of sequence, the regions of repeat density are clearly visible. Numerous repeats were found spanning the entire length of the *hsr* locus, with regions of low density of repeats (*hsr* central and β domains, *orf2* gene) and high density of repeats (USR, DSR and *hsr* variable region).

Figure 4.1 Overview of the *hsr* locus and genetic elements.

A. *H. mustelae* 4298 genome organisation showing the *hsr* locus (HSRL, thick black line) with respect to flanking genes (open arrows, orientation shown as arrow direction). Regions where the sequence of single or both strands of DNA were determined are labelled above the representative line of the strain 4298 genomic DNA. **B**. Enlargement of the HSRL showing features. Repeat sequences are shown as thick black boxes relative to the backbone HSRL DNA line. USR and DSR regions are indicated and DNA functions suggested. The labelled *hsr* gene, and gene component regions (variable, central and β regions) are shown overlaying the repeat sequence blocks. Significant coordinates are indicated. **C**. Suggested protein functions for the Orf2 and Hsr protein. **D**. Consequence of knockout mutagenesis of the *orf2* and *hsr* gene in *H. mustelae* strain 4298.


A.
DNA sequenced prior to this study
DNA sequenced during this study
gyrA hp0702 fhgR
orf2
hsr
glr
HSRL (14919 bp)
SUGGESTED DNA FUNCTIONS
USR & DSR
- reservoir pool of DNA for creating variability in the variable region of the hsr gene
B.
Central domain
Variable region
β domain
USR
DSR
1
7482
8948
11016
12039
14919
orf2
(5858-6445)
hsr
(7482-12039)
C.
SUGGESTED PROTEIN FUNCTIONS
Orf2 protein
- LolA homologue, by amino acid similarity involved in the release and specific transport of lipoproteins from the plasma membrane to the outer membrane (by similarity of function with other LolA proteins)
Hsr protein
persistence in ferret host molecular channel?
DOMAINS:
Variable domain
- host immune evasion?
Central domain
- ring structure?
Autotransporter domain
Hsr membrane anchor
ring assembly?
D.
Knock out mutagenesis
H. MUSTELAE STRAIN 4298 KNOCKOUT CLONE FEATURES
orf2::aphA3'
- non lethal
- no detectable effect on Hsr production
hsr::aphA3'
- non lethal
- absence of 150 kDa protein
- absence of ring structures

Most of the *hsr*-like repeats were inverted with respect to the variable region of the *hsr* gene. However, a cluster of 10 direct repeats (12 – 37 bp) were located directly 3' of the *hsr* gene, representing HSRL coordinates 12041 – 12937 (172 / 933 bp) and are the only directly repeated *hsr*-like segments (with regards to the *hsr*) observed in the entire *hsr* locus. One was a perfect 22 bp inverted repeat (HSRL 12571 – 12592), while the rest ranged in size from 12 – 37 bp. The biological consequences of these or any other repeat sequences were not determined.

Twelve tandemly repeated sequences (6 – 18 bp repeat unit) were located in the HSRL, up to 5 repeat units long (Table 3.12). Some of these (1 – 6 nucleotide unit length) represent examples of short sequence repeats (SSRs) or short tandem repeats (STRs) (van Belkum *et al.*, 1997). Two STRs were mapped to the sequenced repeat region (V4 and V5, Fig 3.6 and Appendix 4a). One of the two STRs located in the V5 region (STR#2) exhibited variable number between *H. mustelae* strains. The generation of these STR is thought to come about either by slipped strand mismatches with a deficient mismatch repair system, or by recombination between mutiple loci containing homologous repeat stretches (van Belkum, *et al.*, 1998). These tandem repeats may represent hotspots for chromosomal gene conversions, as has been suggested for *H. pylori* (Marshall *et al.*, 1998) and like that observed for the gonoccocal pilin variation genes (Howell-Adams and Seifert, 2000).

The 1.5 kb non perfect repeat in the USR of the *hsr* locus gives insights into some interesting features of the repeat sequences. It is possible that gene conversion may have generated a second copy of this 1.5 kb region, although there is a 123 bp gap between the sequences that may have resulted from subsequent recombination events. Alternatively, duplication may have occurred by transformation-mediated homologous recombination with donor DNA from another *H. mustelae* strain, followed by rearrangements in this area, accounting for the variant sequences observed between the repeats. The variant sequences within these repeats may have been produced by successive segmental DNA conversion events, which is supported by the presence of tandem repeats in this region.

4.2.1.2 Repeat features with possible roles in hsr expression

Sequences (REP-like) encoding potential AT-rich hairpin structure(s) (Newbury *et al.*, 1987a)) were identified directly upstream of the *hsr* (IR5) and *orf2* (IR10) genes in transcribed regions (Section 3.4.2.3). In both cases these sequences appeared in the 5' untranslated regions of their respective mRNAs, directly upstream of the putative ribosome binding sites. Two similar inverted repeats were previously noted downstream of the *hsr* gene and indicated as possible terminator structures (O'Toole *et al.*, 1994). Inverted repeats were not detected 3' of the *orf2* gene. Hairpin formation at the 5' and 3' end of mRNA have been implicated in blocking 3' – 5' exoribonuclease digestion, hence stabilising mRNA for translation to proceed (Newbury *et al*, 1987a). Deletion of REP sequences within the *malEFG* operon of *E. coli* resulted in reduced protein production and suggested a role of mRNA stability in controlling gene expression (Newbury *et al.*, 1987b). Some genes may require additional factors involved in stem-loop binding to further stall their mRNA decay (McLaren *et al.*, 1991). This mechanism may also be employed by *H. mustelae* for the abundant production of Hsr from stable mRNAs as shown in Figure 4.2. REP sequences have also been implicated in chromosomal rearrangements in prokaryotes (Stern *et al*, 1984). Interestingly, an almost perfect palindromic 9 bp tandem repeat (2 times) had replaced the native REP-like repeat upstream of the *hsr* gene in the original λE2 derived DNA sequence (Section 1.5.1). However, the mechanism of this DNA arrangement is not known since the source of the 5' sequence beyond the polylinker sequence of the scrambled clone was unknown.

Three significant single nucleotide homopolymeric tracts (>8 nucleotides) were found in the HSRL, directly upstream of the *hsr* gene (T_9, T_{10} and C_{10}). T_{10} was located between the transcription start site (Section 1.3.3) and translational start site and may be implicated in mRNA stabilisation as mentioned in the previous paragraph. T_9 was located approximately 20 bp upstream of transcription start, and was included in part of a weakly conforming (~50%) –35 σ^{70} recognition site (Harvey and Reynolds, 1987; Fig 4.2). T_9 also comprised most of the promoter-proximal half of the inverted repeat IR2. The distal half of IR2 was recognised as part of a putative UP element (Ross, *et al.*, 1993), containing 19/22 matching residues to the most active consensus sequence, -59-

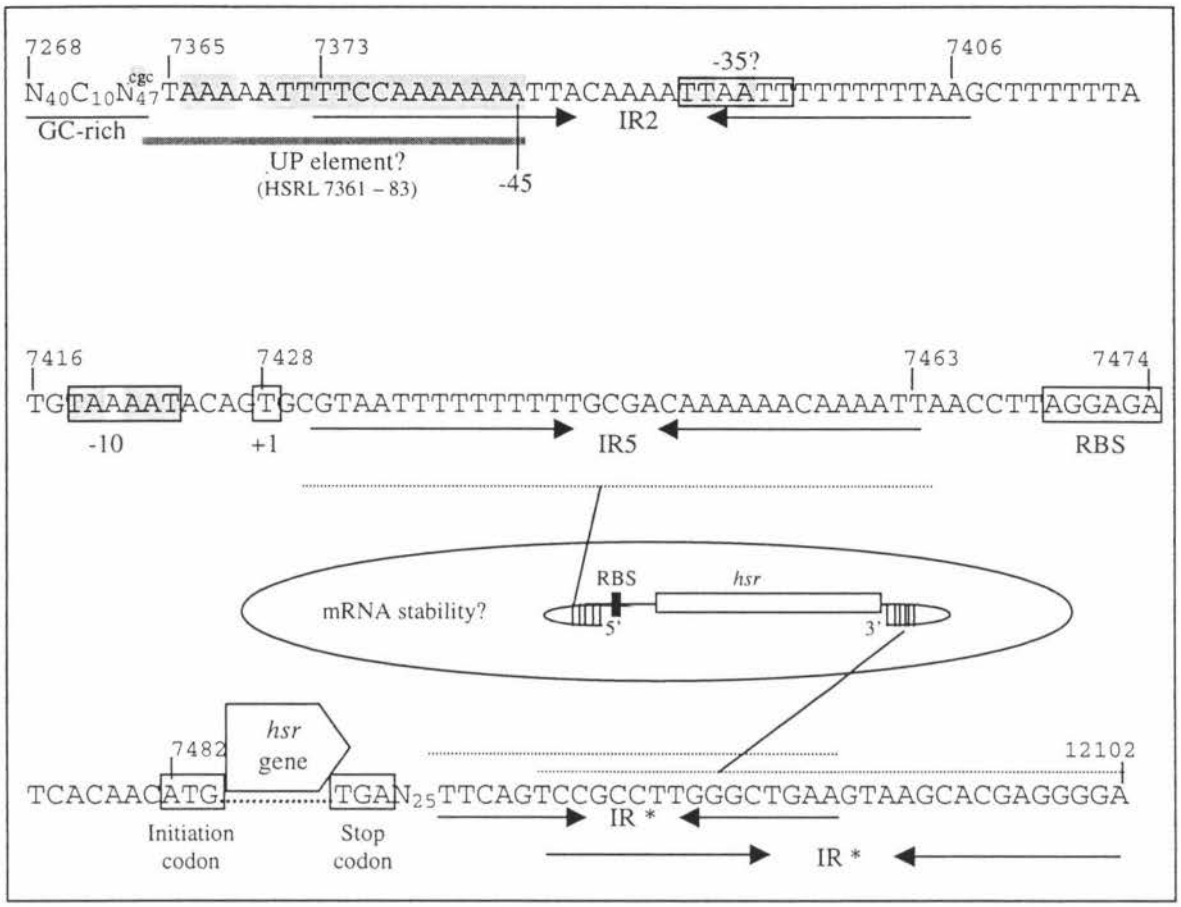


Figure 4.2 Potential regulatory elements of the *hsr* gene

The DNA sequence of the relevant upstream and downstream elements flanking the *hsr* gene is shown, in an abbreviated form. The subscripted number indicates the number of the particular residue directly to its left, and N denotes the DNA sequence of the region abbreviated. Sequence of particular interest is indicated and labelled. Opposing arrows represent imperfect inverted repeats. Inverted repeats IR2 and IR5 were detected in this study, while inverted repeats identified in previous work at annotated as IR* (O'Toole *et al.*, 1994). Potential mRNA hairpin structures are suggested and implicated residues are indicated. UP element (grey bar) extends into the GC-rich region, with relevant nucleotides of the UP element shown in lower case. Shaded nucleotides are conserved in the accepted consensus sequences for the respective elements.

nnAAA(A/T)(A/T) T(A/T)TTTTnnAAAAnnn-38, with respect to the transcription start sequence (Estrem, *et al.*, 1998). The minor groove of each half site of this consensus sequence has been suggested to interact with one each of the two C-terminal domains of the alpha subunit of the RNA polymerase core enzyme, which has been implicated in preferential binding to AT-rich bent DNA (Yasuno *et al.*, 2001). This interaction has been demonstrated to activate gene transcription by a factor of 30-fold (Ross *et al.*, 1993). Additional transcription activator protein(s), binding nearby DNA target binding sequences, may further enhance transcription. For example, the FIS transcription activator protein can increase transcription of the *E. coli* rRNA promoter *rrnB* P1 by a factor of 10, which results in a total 300-fold increase in gene expression (Gaal *et al.*, 1994). However, the positioning of the presumptive UP element for *hsr* does not conform to the consensus spacing, with respect to the transcription start site, spanning -67 to -45 rather than -59 to -38. A "TG" motif is also present immediately upstream of the -10 sequence, but not the consensus "TGn" described for the extended -10 promoter sequence of other bacteria (Kumar *et al.*, 1993). These differences in element spacing and composition may reflect slight differences in the modular assembly of the *H. mustelae* RNA polymerase core enzyme and consequential binding to its target recognition sequences.

The third homopolymeric tract $(C)_{10}$ is located further upstream of the transcription start site (HSRL 7268 - 7364, Fig 4.2). The significance of this sequence is not apparent. However, a region of 97 bp housing this C_{10} tract was extremely GC rich at 66% (HSRL coordinates 7268 – 7364), with 40% (39 / 97) cytosine residues, in contrast to the directly adjacent AT-rich promoter region (80% AT-rich with several A- and T- tracts) and remainder of the HSRL (~55% GC). Methylation of cytosine residues have been implicated in curving of local DNA structure as well as in modulation of gene expression by DNA-binding transcription regulators (Hodges-Garcia and Hagermann, 1992). The GC-rich region, in particular the C_{10} sequence, may be important in recruiting proteins involved in a DNA bending (e.g. integration host factor-like protein, IHF) perhaps to facilitate inversion recombination between CVC motifs in the *hsr* gene and flanking sequence.

Reversible switching of methylation states of the GC-rich region may be involved in the observed phase variation or attenuation of *hsr*, similar to those described for the phase-

variable outer membrane protein of *E. coli*, the antigen 43 autotransporter protein (Henderson *et al.*, 1997). The GC-rich region and the phase variation control elements of Ag43 are positioned similarly to their respective coding sequences. However, the GC-rich sequence in the *hsr* locus is likely to be target for a *H. mustelae* specific cytosine methytransferase enzyme with as yet unknown target sequence specificity. The GC-rich regions may additionally be involved in the mechanism that controls how much Hsr protein is generated. An investigation to identify regions upstream of *hsr*, which are important to *hsr* expression are required to assess these possibilities further.

4.2.2 Distribution of *hsr* repeat sequences within the *hsr* gene

Within the *hsr* gene, the repeats were shown to be confined mainly to the first 2366 bp of the gene, the repeat (variable) region (156 repeats). Nine small (<16 bp) repeat sequences were additionally present in the central domain and one in the β domain, and two 13 bp H group repeats were included in the non variable regions. The cut off size for significant repeat size (12 bp) was the lowest number possible using the DNA Strider program. However, important smaller repeats may have been overlooked. The Neisserial DNA uptake sequence is only 10 bp in length, yet is repeated almost 2000 times in the genomes of *N. meningitidis* (Parkhill *et al.*, 2000b, Section 1.5.1). Similarly, the *H. influenzae* genome carries ~1500 copies of a 9 bp DNA uptake sequence (Gilsdorf, 1998). However, these sequences have not been identified in the genome of *H. pylori* (Marshall *et al.*, 1998). Access to more powerful repeat analysis tools like those used for genome analysis would greatly help to disclose these important smaller sequence repeats, if they are present. Similarly some of the repeat sequences may be coincidental, particularly those that are small and are repeated only twice in the locus. Examination of Figure 4.1 (panel B) shows two areas of clustering of small repeats in the central domain, the most obvious being in the C-terminus of the central domain. These regions cover an area of 243 bp (HSRL coordinates 10770 – 11012) and 194 bp (HSRL coordinates 10018 – 10211), with contribution to repetitive DNA only 63 and 49 bp, respectively. The repeats map is condensed, and repeats appear closer together than they physically are, with respect to the HSRL (Fig. 4.1, panel B). The corresponding regions in the nine *H. mustelae* strains tested showed almost 100% conservation with respect to the 4298 sequence (Appendix 4b). It is tempting to

disregard these apparent clusters, yet the occurrence of highly repetitive DNA repeats (up to 4 times elsewhere in the HSRL), indicates that there is potential for sequence variation. Interestingly, the C-terminus of the passenger domain has been implicated as the hydrophilic portion of the translocation unit (the linker region) in other autotransporter proteins (Maurer *et al.*, 1999). Theoretically this region, in *hsr*, could become exposed transiently, during the translocation process, to host immune selection. Indeed the sequence of the Hsr linker region is significantly different to other autotransporters (not shown), with few conserved amino acids. However, it is also very hydrophilic consistent with the translocation function of the linker region (Section 1.3.2). Investigation of conserved nucleotides in this distinctive *H. mustelae* autotransporter may indicate further important functional amino acids involved in the translocation process to the outer membrane in autotransporter proteins.

4.3 Hsr variability

4.3.1 Variability of the *hsr* gene of different *H. mustelae* strains

In contrast to the strong conservation observed for the central and β regions (>98%) the variable repeat region exhibited considerable variability, with 51.8 – 62.1% identity with the reference strain 4298 (Section 3.2.4.2). Conserved stretches of sequence homology were interrupted by stretches of variable sequence, and these conserved-variable-conserved motifs (CVCs) were identified in protein sequence alignments of the translated partial variable region sequence of all *H. mustelae* isolates (Fig. 3.6).

The variable sequence stretches of the CVCs were almost never repeated, unlike the conserved components of the CVCs. The only exception is the tandem repeat containing region V5 (Section 3.2.4.3). Tandem repeat sequences within variable stretches of repeat region DNA sequence suggest gene conversion or slipped strand mismatch (SSM) as mechanisms for the generation of variant sequences. That these mechanisms might be employed, is supported by the presence of this repeat in the USR and DSR, with 35 copies, and often in short tandem arrays (7 times, comprising 2 – 4 units). Other variant stretches present in several distant places in HSRL suggest non-reciprocal exchange between repeats, or effective duplication (Fig. 4.3A and 4.3Bii).

For example, sequences at variant positions of the respective *hsr* genes that are found also in their flanking sequences (e.g. that coding for Hsr region V5: GAGNANA in strain 4298, or DANAKN of strain F6) could have arisen by either mechanism illustrated in Fig. 4.3A or Fig. 4.3Bii. However, the sequenced F6 *hsr* flank DNA was relatively conserved, with respect to strain 4298 (~95%, Appendix 4e), arguing in favour of gene conversion.

In contrast, examples of Hsr sequences found in strain 4298 at variant positions, which are not underlined in Figure 3.6, could not be found in the translated flanking sequences. Reciprocal exchange of a single copy of relevant encoding DNA sequence, from the *hsr* flanking sequence into the structural gene, may have occurred in these positions. Alternatively, these sequences may have been already lost or replaced. Loss of sequence is a consequence observed for a gene conversion mechanism. Replacement of sequence by reciprocal exchange with DNA from autolysed cells by natural transformation (Section 1.5.3) may have introduced more diversity to the HSRL sequence.

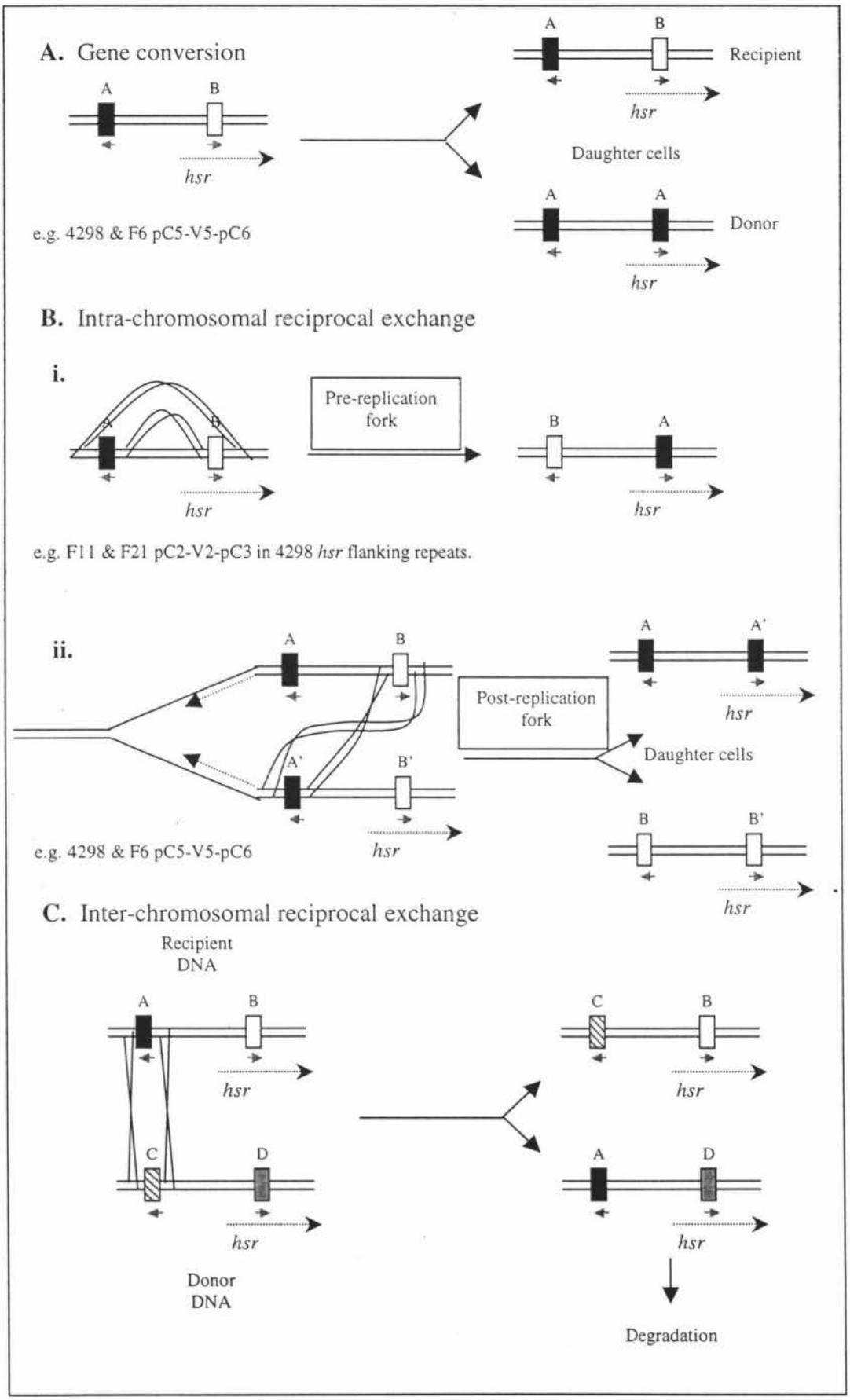
Some DNA sequences found in the *hsr* gene of ferret isolates were also found in the flanking sequence of strain 4298 *hsr* gene. For example, most of the DNA encoding the C2-V2-C3 motif in strains F15 and F21 is found in the strain 4298 flanking sequences, but not in strain 4298 *hsr* gene sequence. These variant sequences may arise by any mechanism illustrated in Figure 4.3. Generation of sequence diversity may also result via intra-chromosomal recombination between flanking repeat sequences, since many homologous repeat tracts are present in the HSRL.

The gonoccocal *pilS* sequences between strains are variable, and are able to recombine with *pilE* several times, resulting in mosaic sequences in *pilE* and antigenic variation of a single strain within a human experimental model (Hamrick *et al*, 2001). Similarly, in *H. mustelae,* chimeric *hsr* genes may result from multiple recombination and gene conversion events. This remains to be substantiated by future work..

The variable regions of the *hsr* genes of *H. mustelae* strains exhibit a high level of DNA diversity, even in those strains isolated from ferrets living in close proximity. These strains have been shown to be most closely related by rRNA gene analysis and *hsr*

Figure 4.3 Potential mechanisms for generation of diversity in the *hsr* locus

Duplicative transposition (Gene conversion) results in original gene segment loss from one daughter cell, replacing it with a copy of gene flanking sequence. **B.** Intra-HSRL reciprocal exchange of gene segments within the *hsr* locus (Pre- and Post-replication fork). **C.** Inter-strain HSRL homologous recombination between HSRL DNA and homologous donor DNA derived from autolysed cells. Participating gene segments (i.e., CVCs) are shown as labelled boxes (A – D). The *hsr* gene and its orientation are represented by the indicated arrows. Short gray arrows depict orientation of the homologous flanking variant sequences, with respect to the *hsr* gene.


A. Gene conversion
A
B
hsr
Recipient
Daughter cells
Donor
e.g. 4298 & F6 pC5-V5-pC6
B. Intra-chromosomal reciprocal exchange
i.
Pre-replication fork
e.g. F11 & F21 pC2-V2-pC3 in 4298 hsr flanking repeats.
ii.
A'
B'
Post-replication fork
Daughter cells
e.g. 4298 & F6 pC5-V5-pC6
C. Inter-chromosomal reciprocal exchange
Recipient DNA
C
D
Donor DNA
Degradation

central and β region sequence comparisons (Section 3.2.2.1 and Section 3.2.4.2). This suggests that *hsr* diversity might have been generated within a single population of cells (or single cell). However, only one *H. mustelae* colony was isolated from each animal in this study. A larger sample size would be required to determine whether variation could occur in the same animal. Another limitation is that there is only one complete HSRL sequence for analysis. Sequence analysis of the *hsr* from several *H. mustelae* colonies from a single animal would be required to determine the frequency and possibly the potential mechanisms resulting in the sequence variation observed between strains.

4.3.2 Serological analysis of Hsr protein variability of *H. mustelae* strains

Western analysis of the reactivity of the antiserum, raised to purified Hsr protein from strain 4298, with Hsr from seven different strains showed different levels of reactivity (Section 3.2.2.2). These results are consistent with consequent ELISA and slot blot analyses of cell surface antisera reactivity of three strains (F15, F21, and 4298) (Forester *et al.*, 2001). Taken singly, the results of this study show protein variability, but do not show differences in surface epitopes. The variability of the different Hsr proteins (Section 3.2) suggests that the variable regions are exposed to the surface via the variable stretches, (i.e., the V sequences of the CVCs, Section 3.2.4.3). These are hydrophilic and hence probably exposed, while the C sequences are more hydrophobic and most likely engaged in interactions with the neighbouring subunits of the ring structure, or with the lipid bilayer. This is in agreement with the predicted N-terminal (passenger domain) surface exposure of autotransporters proteins (Henderson *et al.*, 1998). Insertion and folding in a ring formation of Hsr in the cell envelope is predicted to mask hydrophobic conserved regions of the variable region while exposing the relatively hydrophilic variable regions. F21 antiserum showed lower reactivity to Hsr isolated from host animal than to Hsr from bacteria isolated from two other ferrets (F15 and strain 4298) (Forester *et al.*, 2000). This contrasts with reactions observed with Hsr raised to purified proteins, which reacted best with Hsr protein from same strain. The reason for these results are presumably due to exposure of hidden Hsr epitopes during rabbit immune sera production. Dominant bacterial surface structures are subject to natural immune selection, and an antibody response to Hsr in ferrets has been observed

(Forester *et al.*, 2000). It is therefore reasonable to assume a contribution of the variable amino-terminal epitopes of Hsr to evasion of the host immune response.

4.3.3 Supporting evidence for antigenic variation

Observations made in this study have contributed evidence towards antigenic variation as a mechanism of creating genetic diversity within the *hsr* locus. The 150 kDa Hsr protein from different ferret isolates exhibited size variation, DNA sequence variation (variable repeat region) and variable reactivity to Hsr antisera. RFLP patterns for different strains also differed, more so in strains isolated from geographically distant locations (Section 3.2.3). The *hsr* gene is flanked by hundreds of *hsr*-related sequences, which may serve as a reservoir of genetic variability, in a similar manner as that described for the silent loci of the *N. gonoccocal* pilin (Section 3.4; Gibbs *et al.*, 1989). However, antigenic variants derived from a single strain have not been isolated experimentally. Long-term colonisation of ferrets with a single strain of *H. mustelae* is required to describe the mechanism, which results in Hsr protein variability and antigenic variation.

4.4 Genomic organisation flanking the HSRL

A total of ~22 kb of continuous DNA sequence of the *H. mustelae* strain 4298 gènome has been sequenced, ~2.9 kb of which was done during this study. The genomic organisation around the *H. mustelae hsr* locus shows some similarities to that observed at the corresponding map location in *H. pylori* strain 26695 (Fig. 4.4). These are the *gyrA-hp0702-flgR* gene group and the *glr* gene homologues, which are represented in both genomes. However, the HSRL itself is not present in *H. pylori* (O'Toole *et al.*, 1994). Furthermore, the gene homologues, in *H. pylori*, are arranged differently to *H. mustelae*, with the *glr* gene found approximately 165 kb away from the *gyrA-hp0702-flgR* gene group in the chromosome strain 26695 (Fig 4.4).

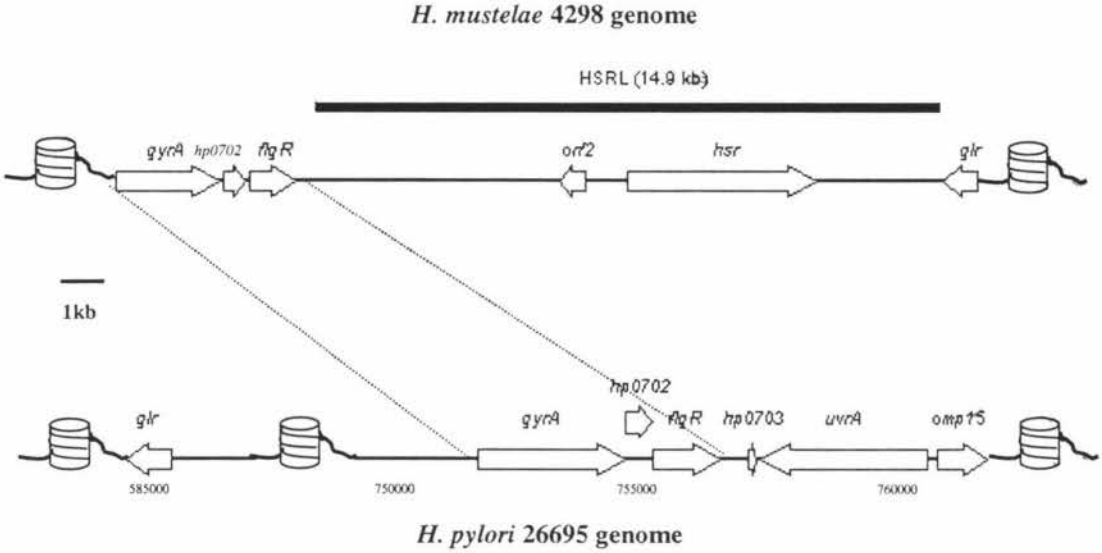


Figure 4.4 Comparison of the genetic organisation around the *H. mustelae* HSRL with respect to the corresponding *H. pylori* strain 26695 DNA and protein homologues.

The putative genes flanking the HSRL of *H. mustelae* strain 4289 are shown relative to the neighbouring genetic loci. The organisation of the corresponding region from *H. pylori* strain 26695 is displayed for comparison. The *hsr* locus is portrayed as a thick black line (labelled). Genes are depicted by open arrows. Actual genomic coordinates are given for *H. pylori* strain 26695. Distances (kb) are indicated for HSRL flanking genes (i.e., bar line equals 1 kb of *H. mustelae* chromosomal DNA). Cylindrical shaped objects, flanking indicated genes, represents non-relevant DNA. Dotted lines connect the putative genes found in the sequence upstream of the *hsr* locus of *H. mustelae* strain 4298 with the corresponding gene group of *H. pylori* strain 26695.

4.5 Orf2 is related to the LolA family of lipoprotein carrier proteins.

BLAST protein database searches revealed similarity of the hypothetical protein Orf2 in the HSRL of *H. mustelae* strain 4298 to the LolA family of periplasmic lipoprotein carrier molecules. Forced alignment of the Orf2 sequence with the *H. pylori* LolA proteins (Fig. 3.11) convincingly showed significant regions of homology, with all three proteins sharing 34.1% identity over the whole molecule, and as high as 58.6% identity over a stretch of 58 amino acid residues. Aside from *H. pylori*, several other LolA proteins of other bacteria were also identified as significantly similar, by BLAST searches, but with lower identity values. Those were LolA proteins of *C. jejuni* (AL139076), *P. aeruginosa* (AAG06002), *C. burnetii* (X75627), and lastly *N. meningitidis* (AL162754), which had the next highest identity values after the *H. pylori* proteins with the Orf2 protein. All of these yielded values were no lower than 23% over 176 amino acids. By protein sequence similarity, the Orf2 protein could be classified as the *H. mustelae* LolA protein homologue. The LolA protein is essential for the release and specific transport of *E. coli* lipoproteins from the periplasmic side of the inner membrane to the periplasmic side of the outer membrane (Matsuyama *et al.*, 1995; Tajima *et al.*, 1998). This process requires: energy from ATP (Yakushi *et al.*, 1998), the accessory proteins LolB (Matsuyama *et al.*, 1997; Yokota *et al.*, 1999), and the ABC transporter complex, LolCDE (Yakushi *et al*, 2000). The sorting signal at position +2 of the mature lipoprotein specifies membrane destination and is required for correct localisation (Seydel *et al*, 1999). Protein homologues for some of these accessory proteins are found in both of the sequenced *H. pylori* strains (26695 and J99) and several other Gram-negative bacteria. It is therefore possible that a similar system is employed by *H. mustelae* for lipoprotein localisation.

The upstream promoter elements of the *orf2* gene were not determined experimentally. Putative $^{-}$10 and $^{-}$35 sequences were identified by visual screening in the *orf2* upstream sequence (Section 3.5.1.1). A 5' untranslated hairpin was possible for mRNA structure (Section 3.4.2.3). This structure may stall mRNA decay from the 5' side, though a similar loop structure was not detected in the 3' non-translated region of *orf2*. Furthermore, the detection of the Orf2 product was prevented by the lack of specific antibody.

Orf2 is apparently produced with a 17 amino acid signal peptide (section 3.6.1.2), suggesting it is an exported protein. By size and content, the putative Orf2 signal sequence is similar to other LolA proteins. In contrast, the putative mature protein is 12 amino acids larger than the *H. pylori* LolA proteins. The 14 amino acid polar/charged overhang at the N-terminus of the putative mature Orf2 protein (Orf2 mature protein coordinates 2-9, Fig. 3.11) exhibits an overall positive charge, which contributes to the hydrophilic nature observed for the protein (GRAVY score -0.293). These residues potentially form ionic bonds with negatively charged cell components, for example, negatively charged phosphate groups in the outer phospholipid leaflet, which are exposed in the periplasm. Alternatively this block of amino acids may determine the specificity of Orf2 binding protein to proteins other than lipoproteins, but this has yet to be experimentally proven.

Knock out mutagenesis of the *orf2* gene did not appear to have any negative effect on the production of the Hsr protein in the Hsr expressing strain of *H. mustelae*. The preliminary results of recent transformation-based experiments with the Orf2 knockout of the non-expressing strain have implicated Orf2 in the reversion to Hsr expression of the non-expressing variant. During construction of an Orf2 knockout, in an attenuated Hsr$^-$ strain 4298, approximately 20% of the clones assessed for Hsr expression had reverted to the Hsr-expressing phenotype. The Orf2+ strain, with a deletion of a neighbouring gene, *flgR* (Fig. 4.1) did not produce revertants (O'Toole and Forester, unpublished results). Orf2 is predicted to be present in the periplasm (and therefore would be presumed to exert functional effect at the protein level). Yet loss of Orf2 appears to be involved in reversion of attenuation. However, the mechanism of Hsr attenuation / reversion is unknown and requires further investigation to identify the components involved, together with their modes of interactions with each other.

The LolA protein in *H. pylori* strain 26695 genome (accession AE000511), is surrounded by hypothetical proteins, and does not resemble the genomic organisation flanking observed for the *hsr* locus (Fig 4.1). The precise relationship between Orf2 and Hsr with respect to the positioning of the *orf2* gene within the *hsr* locus needs clarification. A single repeat occurred within *orf2*. This repeat (#314) was duplicated in the DSR region, which is most likely coincidental given that the repeat is small (12

bp) and occurs only twice in 15 kb of DNA sequence.

4.6 Function of the Hsr protein

Production of a large and highly expressed protein must be metabolically expensive to the cell. Therefore, it is reasonable to assume that the Hsr protein must be providing an important function to be maintained in all *H. mustelae* primary cultures isolated from naturally infected ferrets isolated and examined for Hsr thus far. Knockout mutants of Hsr are stably maintained *in vitro* (Section 3.5.3), indicating Hsr probably does not possess essential housekeeping functions. Furthermore, Hsr attenuation has been observed in serially-passaged cultures of strain 4298 (O'Toole *et al.*, 1994).

Hsr belongs to the autotransporter (AT) family of proteins. The function of the AT β-domain and leader sequences are well established (Section 1.3.2). However, the pathogenic function of the passenger domain (variable and central regions) still remains unsolved. Hsr is not an adhesin (Forester *et al.*, 2001), nor a haemagglutinin, and may not possess anti-phagocytic properties (O'Toole *et al.*, 1994). Hsr lacks known serine protease active sites, and therefore is not a serine protease. Hsr is retained on the outer membrane, and its release has not been detected in the laboratory, thus it is unlikely to be a toxin. Hsr is well placed to contribute to immune evasion. The preliminary results of recent experimental infections of ferrets with Hsr knockout mutants have suggested a role in establishing long term infection of ferrets with *H. mustelae* (i.e., persistence) (Fox *et al*, unpublished results). Persistence is probably facilitated by antigenic variation, but that has not been proven experimentally (refer to Section 4.4).

Phase variation may contribute to Hsr ring loss in the laboratory. There were no characteristic di-, tetra-, or pentameric tandem repeats implicated in translation control of phase variation (Henderson, *et al.*, 1998). The polyC tract upstream of the *hsr* gene does not appear to vary in the attenuated Hsr⁻ strain (Forester and O'Toole, unpublished results), so it is unlikely to exhibit a volume control type gene expression as observed for the *N. meningitidis* Opc adhesin protein (Sarkari *et al.*, 1994). Non-expression of the *hsr* gene was observed as a switch of phenotype, and not derived from a single colony (Forester and O'Toole, unpublished results). It is possible that Hsr expression is

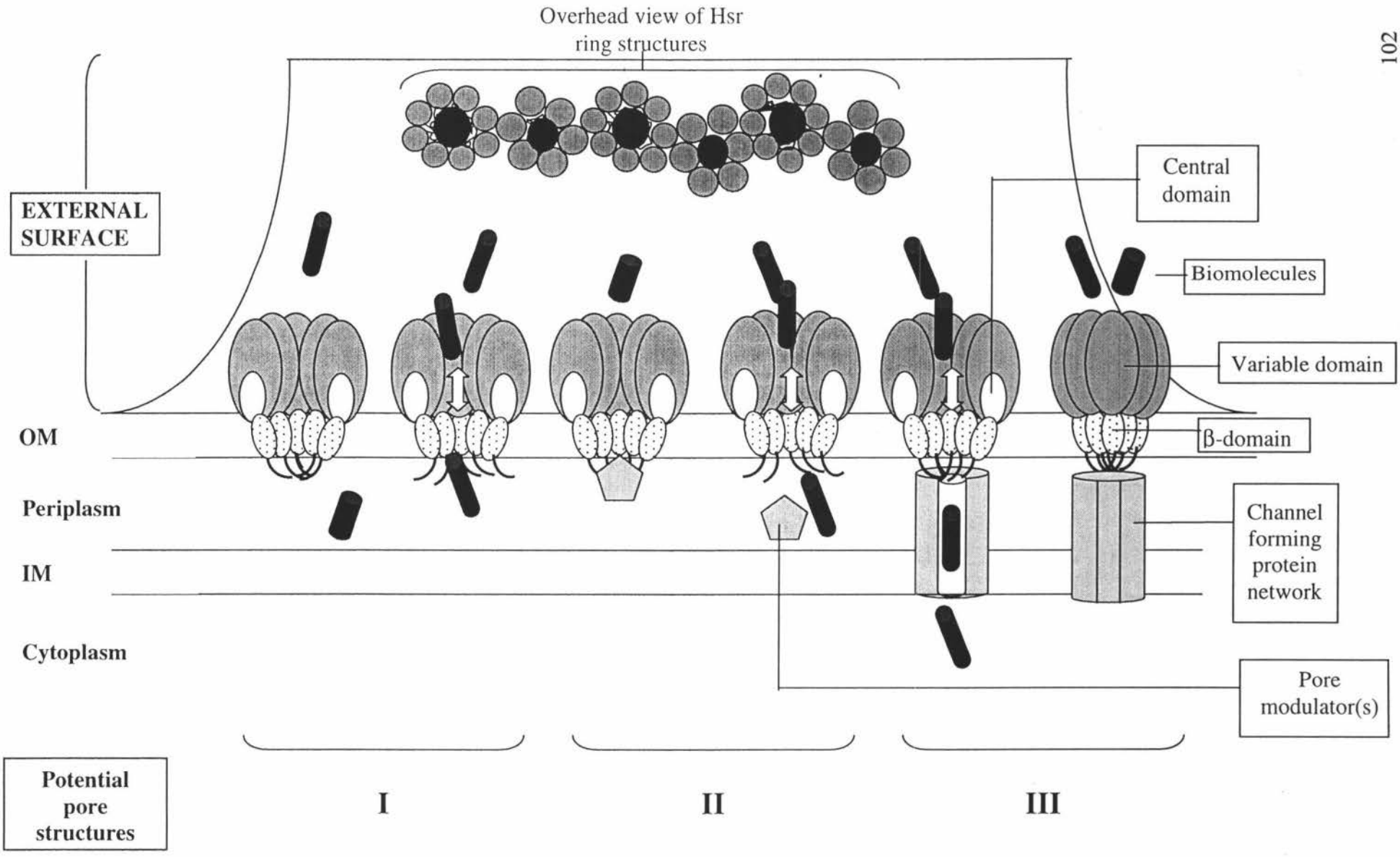
physically repressed during attenuation, as a consequence of some, as yet unknown environmental factors, for example low nutrient supply due to overcrowding.

The ring-shaped appearances of the Hsr may be providing functions other than host immune evasion. Hsr monomers appear to group together and form 4.25 nm rings in the outer membrane upon visualisation at high magnification (O'Toole *et al*, 1994; Forester *et al.*, 2000). Six or seven monomers appear to associate together to form a ring (Fig. 4.5), but a high-resolution electron micrograph showing distinct monomeric protein outlines has not been obtained. Some ring forming proteins, such as YscC, form channels (Koster *et al.*, 1997). Thus, the Hsr multimeric rings could form channels, large enough for small biomolecules (sugars and amino acids) and some smaller macromolecules (nucleic acids, peptides) to enter. In addition, Hsr shares some similarities with S-layer proteins, such as abundant expression, size range (46 – 169 kDa, Boot and Pouwels, 1996), acidic pI, and surface exposure. If Hsr rings are indeed capable of pore formation, then it may perform a "molecular sieve" function similar to that described for S-layer proteins (Sara and Sleytr, 2000) and other pore forming membrane proteins (e.g. porins).

A notable difference between the Hsr and other autotransporter proteins (Section 1.3.2; Henderson *et al.*, 1998) is a stretch of 34 predominantly charged (52.9%) amino acids following the final transmembrane β-sheet signature motif. Predictably, this region (C-terminal charged region) is hydrophilic in nature and probably extends into the periplasm. The sequence of the final transmembrane spanning β-sheet is shown to be an important factor involved in targeting autotransporter proteins to outer membrane and for correct assembly within the outer membrane (Hendrixson *et al.*, 1997; de Cock *et al.*, 1997). In order to perform this function efficiently, looping back of the C-terminally charged domain is predicted to be required. The extended autotransporter signal sequence contains approximately the same proportion of charged residues as does the C-terminal protein segment and is suggested to be involved in the recruitment of accessory proteins in autotransporters (Henderson *et al.*, 1998). Possible roles for the C-terminal peptide may include recruitment of *H. mustelae*-specific periplasmic protein(s) to aid efficient export and assembly of the barrel structure into the outer membrane (e.g. periplasmic chaperone type proteins). Alternatively, the C-terminal peptide could form a binding site for proteins involved in the structure or regulation of

Figure 4.5 Model of potential Hsr structure and function

Schematic views of the *H. mustelae* cell surface curved to show: Top, Hsr ring structures from the outside milieu, looking down on membrane; Bottom, the side view. Membrane compartments of the cell are labelled. The three distinct protein regions and other proteins or molecules proposed to be related to Hsr function are indicated. Three proposed of Hsr ring pore structures are shown. Structure I: Formation of self-gated channel via associations between C-terminal charged regions. Structure II: Channels gated by interactions with regulatory protein. Structure III: Channels formed from external milieu to the cytoplasm of the cell by interaction with a cytoplasmic-membrane associated channel-forming protein network through interactions of C-terminal charged region.


Overhead view of Hsr
ring structures
EXTERNAL
SURFACE
Central
domain
Biomolecules
Variable domain
β-domain
OM
Periplasm
IM
Cytoplasm
Channel
forming
protein
network
Pore
modulator(s)
Potential
pore
structures
I
II
III

macromolecular transport. Interactions of these proteins with the Hsr protein could produce a channel extending from the external surface through to the cell cytoplasm (Fig. 4.5), perhaps similar to the arrangement of proteins forming the T-pilus structure of *Agrobacterium tumerfaciens* (Das and Xie, 2000). Important protein components of this DNA export system (type IV), VirB9 and Vir10, are homologues of ComB2 and ComB3 proteins of *H. pylori*, involved in natural transformation *H. pylori*, suggestive of a similar organisation of DNA uptake machinery in *H. pylori* (Hofreuter *et al.*, 2000). Homologues of the *comB* locus were not detected by hybridisation experiments on the *H. mustelae* genome (Smeets *et al.*, 2001; Hofreuter *et al.*, 1998). Furthermore, successful transformation has occurred in the Hsr-attenuated *H. mustelae* strain, arguing that Hsr is not an essential part of the DNA uptake system.

Interestingly, the β-barrel structure of the Hsr protein showed very little sequence homology to other autotransporter proteins, particularly in the middle section of the β-barrel region (not shown). The β-domain of Hsr may have another function in ring assembly via interactions between β-barrel structures. Other outer membrane proteins are associated together through transmembrane spanning domains (e.g. PhoE porin, de Cock *et al.*, 1997). The presumed Hsr ring pore may also form a gated channel to the periplasm via interactions of the charged C-terminal peptides of individual ring monomers with each other or with positively charged residues of other protein(s). A model of what Hsr might look like based on the results of the DNA sequence data in combination with some of the suggested channel functions is depicted in Figure 4.5.

5.0 SUMMARY OF THE MAIN OUTCOMES OF THIS STUDY AND SUGGESTED FUTURE STUDY DIRECTIONS

5.1 *H. mustelae* and New Zealand Mustelids

This study represents the first successful isolation of *Helicobacter mustelae* from the stomachs of ferrets from two distinct locations in New Zealand. This finding has extended investigations to determining the presence of *H. mustelae* in other Mustelids in New Zealand. Here, ferrets, stoats, and weasels represent major ecological pests. The observation that the *H. mustelae* host range is seemingly limited, and that *H. mustelae* is apparently present in one of the introduced *Mustelidae* species, has generated interest of *H. mustelae* as an antigen delivery system for Mustelid population control via immunocontraception. Hsr, as the major surface exposed protein, is subsequently being scrutinised as a potential candidate for self-antigen exposure.

5.2 Hsr ring structure and suggested functions

The 14919 bp *hsr* locus of strain 4298 contained DNA sequence repeats, which were homologous to the first half of the *hsr* gene. The passenger domain of the proposed Hsr autotransporter has been divided into two domains, the variable repeat region and the conserved central domain, as a consequence of the information obtained from the repeated sequence analysis. The DNA repeat analysis has identified many microcassettes (conserved-variable-conserved domains, CVCs) that may participate in recombination to generate diversity within the variable repeat domain of the Hsr protein. The variant (V) sequences of the CVCs are more hydrophilic than their flanking constant sequences, and thus are likely to be exposed on the surface of the Hsr protein. Based on these outcomes, and previous information regarding the conserved integral C-terminal β-barrel, a structural organisation has been predicted for the domains each Hsr protein monomer relative to the bacterial cell surface (Fig. 4.5). Components involved in ring formation and ring-to-ring associations are yet to be determined. In addition, the possibility of a channel-associated function requires investigation.

5.3 *hsr*-like repeats and antigenic variation

The DNA sequence of the 14919 bp *hsr* locus of strain 4298 was completed and was shown to contain several different repeat types, including, 343 dispersed repeats (12-741 bp long) and 14 perfect tandem repeats (1-18 bp long) along the length of the *hsr* locus. Evidence has been provided in this study supporting antigenic variation of the Hsr protein. A total of ~10 kb of DNA sequence, ~7 kb upstream and ~3 kb downstream of the *hsr* gene, may serve as a reservoir for sequence variation of the *hsr* gene. In addition, conserved-variable-conserved motifs found in the flanking repeat DNA and in the variable repeat region of the *hsr* gene of nine different *H. mustelae* isolates, suggested the possibility of the generation of several potential variants of the Hsr protein. The Hsr proteins from six different *H. mustelae* strains exhibited variability in protein size and reactivity to anti-Hsr antisera. However, antigenic variation is still yet to be demonstrated within a single cell. Long term infection of a ferret with the characterised *H. mustelae* strain 4298 and subsequent sequence analysis, after passaging, may demonstrate an antigenic variation mechanism. To facilitate the detection of DNA arrangements, attempts were made to construct an epitope-tagged Hsr molecule. However, this was made difficult by the lack of phenotypic counter-selection tools for *H. mustelae*. (Forester, unpublished results). Generation of more selection systems would be beneficial to all future work in *H. mustelae*.

5.4 *hsr* expression

Analysis of the repeat sequences in the *hsr* locus has enabled the identification of potential regulatory DNA/mRNA elements upstream of the *hsr* gene that may be involved in the observed abundant expression, in particular the potential UP element and mRNA hairpin/ inverted repeat. Site-directed mutagenesis studies in conjunction with DNA footprinting experiments and serological data may confirm these DNA sequence elements as being involved in abundant *hsr* gene expression. These procedures could also give clues as to a role for the GC-rich region in hsr expression. This region may be implicated in inversion recombination, since 96% of the repeat sequences are inverted with respect to the *hsr* gene.

5.5 Orf2 – the LolA homologue

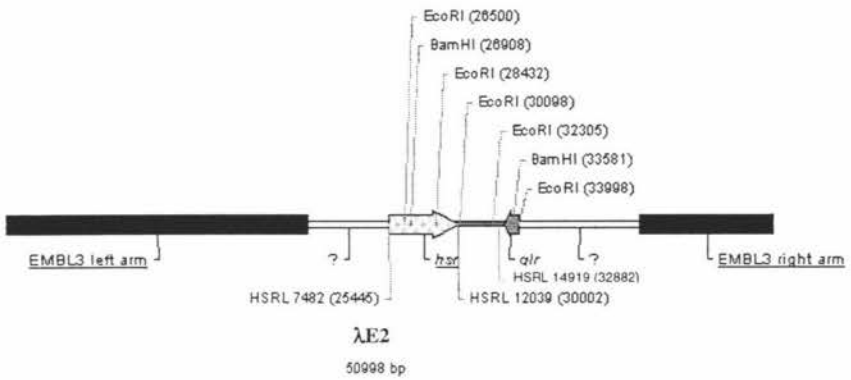
In this study, the putative *orf2* gene was identified and was shown to encode a homologue of LolA, a periplasmic carrier protein involved in selective lipoprotein transport across the periplasm. An N-terminal extension of the predicted mature Orf2 protein was detected upon alignment to *H. pylori* LolA homologues. This extension contains14 charged and polar amino acids, which could form ionic bonds with positively charged cellular components in the periplasm or to alter the specificity of target binding protein to proteins other than lipoproteins. The curious placement of the putative *orf2* gene within the USR region of the *hsr* locus has generated questions regarding a possible relationship between Orf2 and Hsr. Deletion of Orf2, in this study, showed no detectable effect on Hsr production in an Hsr expressing strain. However, subsequent experiments have tentatively implicated Orf2 in the reversion of the Hsr$^-$ attenuated phenotype. Generation of a specific antibody for the detection of LolA would facilitate future attempts to elucidate the function of Orf2 of *H. mustelae* and its N-terminal extension, the target protein of LolA, and the precise nature of the relationship of Orf2 with regard to *hsr* gene expression.

APPENDIX 1

The following maps illustrate sizes, genetic components and their orientation, origins of replication and multiple cloning sites of the vectors used in this study. The maps also show restriction sites and priming sites relevant for cloning procedures. Vectors generated by persons other than the author are referenced.

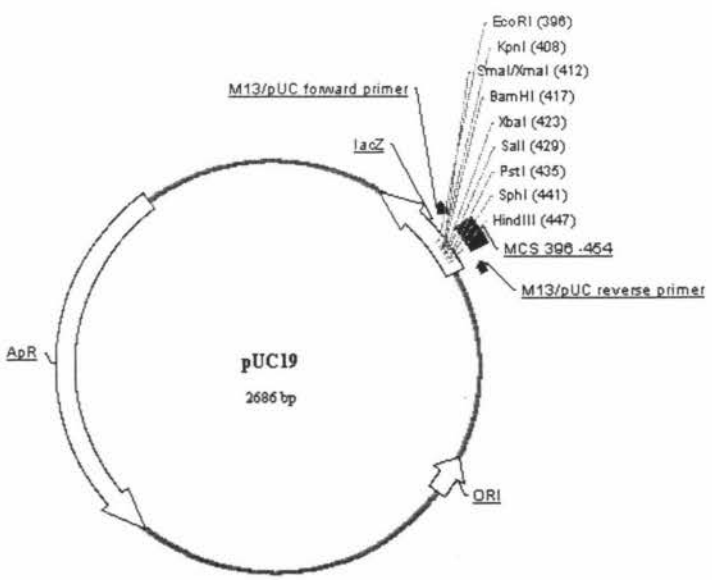
A1.1 Lambda clone λE2

A 22 kb genomic fragment from *H. mustelae 4298* inserted into the EMBL3 replacement vector. Generated by Dr P.W. O'Toole, Massey University, New Zealand (O'Toole *et al.*, 1994). DNA co-ordinates are given in parentheses.



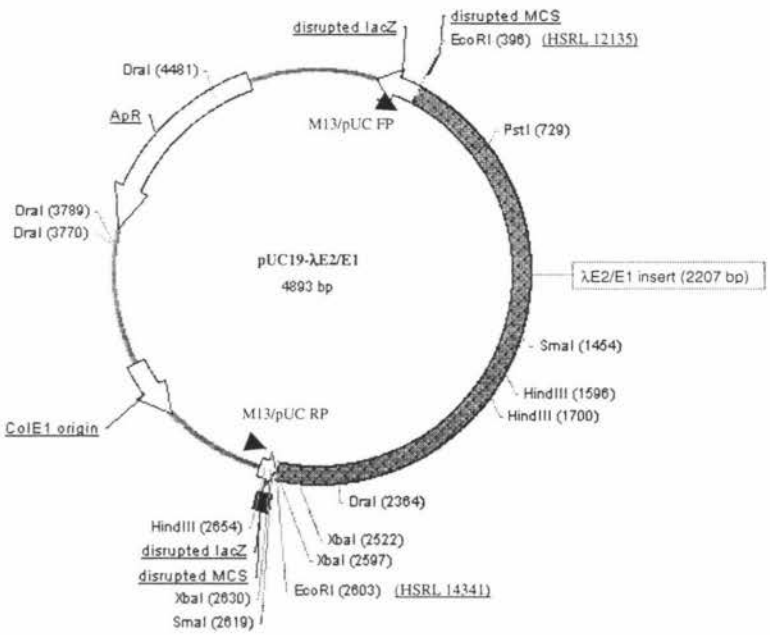
A1.2 Plasmid pUC19

General cloning vector.
GenBank accession X02514
(Yanisch-Perron *et al.*,
1985).



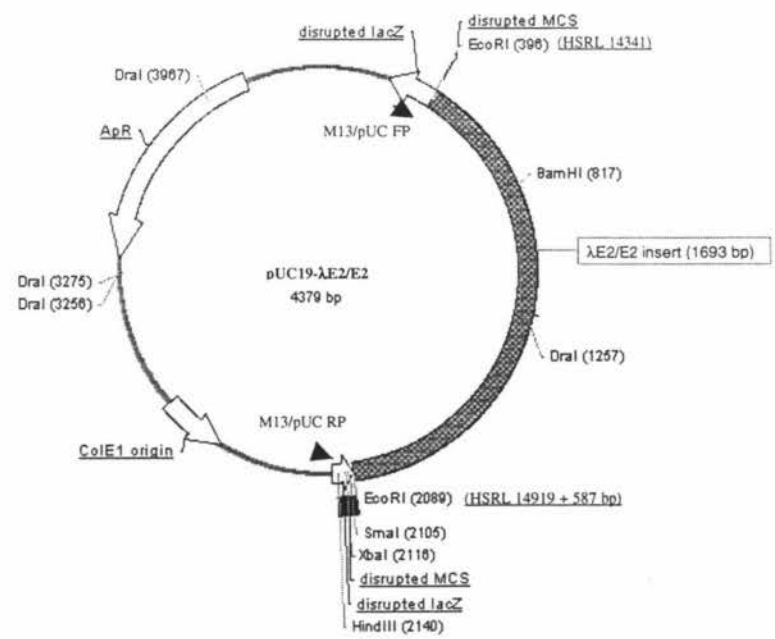
A1.3 Plasmid pUC19-λE2/E1

A 2.2 kb λE2 fragment (strain 4298 genome derived) in the *Eco* RI site of pUC19.



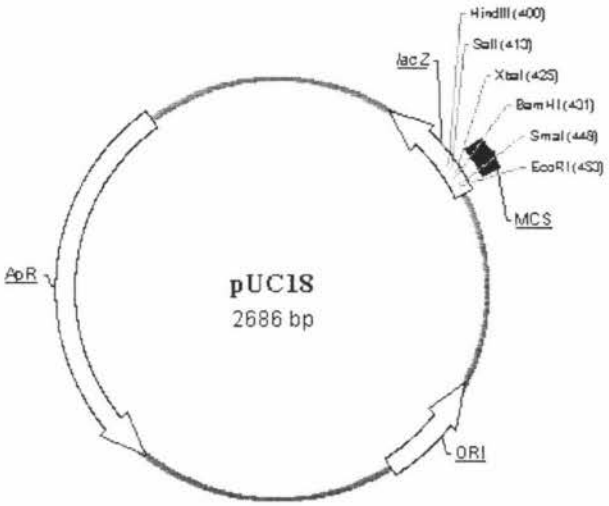
A1.4 Plasmid pUC19-λE2/E2

A 1.7 kb λE2 DNA fragment (4298 genome derived) in the *Eco* RI site of pUC19.



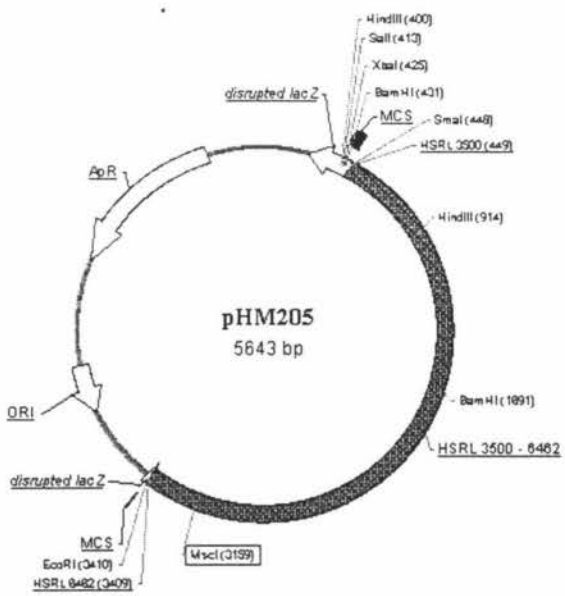
A1.5 Plasmid pUC18

General cloning vector (Yanisch *et al.*, 1985).



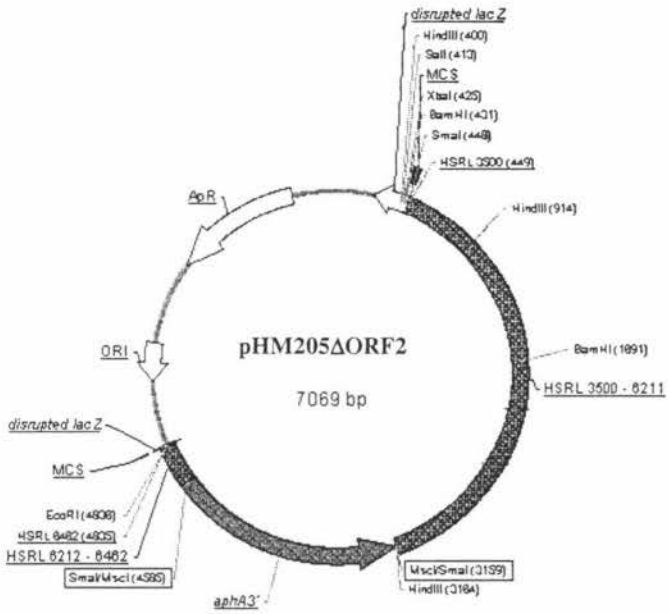
A1.6 Plasmid pHM205

A 2.9 kb HSRL fragment from strain 4298 in *Eco* RI and *Sma* I sites of pUC18.



A1.7 Plasmid pHM205ΔORF2

The *aphA3* gene inserted into the *Msc* I site of plasmid pHM205



APPENDIX 2

The complete DNA sequence of the *hsr* locus. The 14919 bp *hsr* locus is shown. The double stranded nucleotide sequence of the *hsr* locus coordinates 1-12144 was available at the onset of this study (Accession # L15629 and O'Toole and unpublished work). Both strands of the *hsr* locus nucleotide coordinates 12145-14919 were sequenced during this study. Numbers to the left and above the DNA sequence indicate *hsr* locus nucleotide reference coordinates. Open reading frame names are shown inside arrows that mark their orientation. Putative upstream elements are boxed and labelled 5' of their respective open reading frames. Amino acid sequences from the open reading frames are annotated above or below their coding DNA. Restriction enzyme target sites for *Bam* HI, *Cla* I, *Dra* I, *Eco* RI, *Hind* III, *Pvu* II, *Sma* I and *Xba* I are underlined and indicated above the relevant nucleotide sets. Primer sequences are shaded and labelled together with arrows showing the direction of amplification. Primer modifications are illustrated as follows: ✂ , 5' end of the primer includes a restriction enzyme target site; ≈, primer was purified by polyacrylamide gel electrophoresis; Ⓟ, 5' end of primer is phosphorylated. Underlined regions represent an overlap between 2 primers. Double underlined unshaded regions denote mismatched nucleotides in the multi-priming oligonucleotides. This completed sequence was submitted to GenBank (Accession AF254134).

```
1
CCG CCG CTG CCA GGG CTA GAA GTT CCA TTA TTT TGG AAG TAA AGG TTA TTT CCA CCC ACA
GGC GGC GAC GGT CCC GAT CTT CAA GGT AAT AAA ACC TTC ATT TCC AAT AAA GGT GGG TGT
             ⇦ ntf026 ⇦
61
CCC GCA CCT TCT GCT TGA TTT GTT ACT GCA TTT CCT GTG AGC AAG ATA GGA TTA CTT ACG
GGG CGT GGA AGA CGA ACT AAA CAA TGA CGT AAA GGA CAC TCG TTC TAT CCT AAT GAA TGC

121
ATA TTG CTA GTA TTC ATT ACT AGC GTA TTC ACT CCT TGC CCC ATA CTA TTT ACT GCT GCA
TAT AAC GAT CAT AAG TAA TGA TCG CAT AAG TGA GGA ACG GGG TAT GAT AAA TGA CGA CGT

181
TTC ACA AGC CCA CCA TCG GTA AAC ACG AGA TTA TTA GCA CCA TTA TTG TTG CTG TGA TAT
AAG TGT TCG GGT GGT AGC CAT TTG TGC TCT AAT AAT CGT GGT AAT AAC AAC GAC ACT ATA

241            Pvu II
TTG CAT GCT ACA GCT GCA CCT CCA CTT GCT GCA CCT GCA GCA TGC AAA AAA CTC GAC TGT
AAC GTA CGA TGT CGA CGT GGA GGT GAA CGA CGT GGA CGT CGT ACG TTT TTT GAG CTG ACA

301
AAT TCC TTG AAA GTC TTC TTA ATG TAT GAA AAA TAA AGG AGT AGG ATG ATT GCT ATC CAC
TTA AGG AAC TTT CAG AAG AAT TAC ATA CTT TTT ATT TCC TCA TCC TAC TAA CGA TAG GTG

361
AAT GCA AAA GCC AAA ATT TGG GGC AGG AAA AGA GGG CTA AGA GCT AAT ACC CTC CGC AAT
TTA CGT TTT CGG TTT TAA ACC CCG TCC TTT TCT CCC GAT TCT CGA TTA TGG GAG GCG TTA
```

421 *Eco* RI
ACG CGG AGG CTT TAC CCT AGA ATT CTG ATT ATT AAC CAA ATT ATA GCT ATA ATC AGT TGC
TGC GCC TCC GAA ATG GGA TCT TAA GAC TAA TAA TTG GTT TAA TAT CGA TAT TAG TCA ACG

481
AAG AGT AAG ATG ATT AAC TCT TTG TTC ATT TTA TAG CCT CCT TTC TTA ATA AAG TAG GCT
TTC TCA TTC TAC TAA TTG AGA AAC AAG TAA AAT ATC GGA GGA AAG AAT TAT TTC ATC CGA

541
ACC TTG AGA GGT TGC GCC CTC TTA AGG CAA CCT CTT TTA TAA TTA TTC CAT AGA AAC CAT
TGG AAC TCT CCA ACG CGG GAG AAT TCC GTT GGA GAA AAT ATT AAT AAG GTA TCT TTG GTA
⇦ ntf024 ⇦

601
AAC CCC GCT TAT TCT CCT TAA ATA AAT GCT TTT ATT AGC ATT CCC TAT TTC CCA CAT AAG
TTG GGG CGA ATA AGA GGA ATT TAT TTA CGA AAA TAA TCG TAA GGG ATA AAG GGT GTA TTC

661
CAA CCT GAT TTA GCA CTG TTT GAG CAG GAT TAT TGT TAA CAG TTC CAC CGC TAG AGC TGC
GTT GGA CTA AAT CGT GAC AAA CTC GTC CTA ATA ACA ATT GTC AAG GTG GCG ATC TCG ACG

721
CGC CAG GGC TAG AAG TTC CAT TAT TTT AGA GGA TTT GTC TAG TGT TAT AGC GAT TTC ATT
GCG GTC CCG ATC TTC AAG GTA ATA AAA TCT CCT AAA CAG ATC ACA ATA TCG CTA AAG TAA
⇦ ntf026 ⇦

781 ⇨ ntf025 ⇨
CCC CCA CCT CTA CTG TTG CTG GAT CTG CAT TAC CTG CAC CAC CTG CTG CAT TTG CAC CAT
GGG GGT GGA GAT GAC AAC GAC CTA GAC GTA ATG GAC GTG GTG GAC GAC GTA AAC GTG GTA

841
TAA ATG TAA TAT TAT TTG TTC CTG TAC CAC TGG TTC CAT TAG TAT TTC CTG TTA CGA TGT
ATT TAC ATT ATA ATA AAC AAG GAC ATG GTG ACC AAG GTA ATC ATA AAG GAC AAT GCT ACA

901
TTC CAG TCA TGG TAG CAC CAC CAC TAA AAG TAG CAT TGA ATG TTG CAC CGG TAT TGA CAT
AAG GTC AGT ACC ATC GTG GTG GTG ATT TTC ATC GTA ACT TAC AAC GTG GCC ATA ACT GTA

961 ⇨ ntf034 ⇨
TGA TTA CAG GCC CTC CGC TTG CGA TGT TTG TAT TTA GAG TTG CAT TTC CAG CGC CAA GGA
ACT AAT GTC CGG GAG GCG AAC GCT ACA AAC ATA AAT CTC AAC GTA AAG GTC GCG GTT CCT

1021 ⇨ ✂ (*Bgl* II) ntf049
⇨
TGT TGA GAT TCC CAG TTA GTG CAT TAG CAT TAG CAT TTC CAC CAC CAA GAT TCA GGG TGA
ACA ACT CTA AGG GTC AAT CAC GTA ATC GTA ATC GTA AAG GTG GTG GTT CTA AGT CCC ACT
⇦ ntf048 ✂ (*Bgl* II) ⇦

1081
AGG TTT TGG CTG GAT TGC CAG CAA TCC CAC CAA AAT CAA AAT TCA AGT TTA CTG GCT TTC
TCC AAA ACC GAC CTA ACG GTC GTT AGG GTG GTT TTA GTT TTA AGT TCA AAT GAC CGA AAG

1141
CTG TAG CGC CTA GAT TTT GAT TTG CTG GAT CTG CTG GGT CTG CAC CTT GTG CAT TTC CAC
GAC ATC GCG GAT CTA AAA CTA AAC GAC CTA GAC GAC CCA GAC GTG GAA CAC GTA AAG GTG

1201
CTG CAC CTA CTA GCT TGG CAT CCC CGT TGC ATT TGC GCT ATC GTT ATT AAA TTT TTG CAA
GAC GTG GAT GAT CGA ACC GTA GGG GCA ACG TAA ACG CGA TAG CAA TAA TTT AAA AAC GTT
⇦ ntf023 ⇦

1261
ATA ATT TTC TAT CTG CTG GCC TCT TGC TTG AGC ATT AGT ATT TCC ACT TAC ACC ACC AAG
TAT TAA AAG ATA GAC GAC CGG AGA ACG AAC TCG TAA TCA TAA AGG TGA ATG TGG TGG TTC

1321
CCC ACC TGC TTG TTC CAT CAA CTT TCC TAG ATT GCC AGC AAG CCC AGT AGT AAG CCC AAT
GGG TGG ACG AAC AAG GTA GTT GAA AGG ATC TAA CGG TCG TTC GGG TCA TCA TTC GGG TTA

1381
TGA AAA ATA TTT TTG GAT TTG TCG CTG TGA TAT TTG CAG CGG GCG CTG GGG CTG CTT GAT
ACT TTT TAT AAA AAC CTA AAC AGC GAC ACT ATA AAC GTC GCC CGC GAC CCC GAC GAA CTA

1441 *Hind* III *Pvu* II
CTG CAC CTG CTA CTG GAT TTC CTT CAA GCT TGA TTT GTG TCT GCA CCA GCT GGG TTT GTG
GAC GTG GAC GAT GAC CTA AAG GAA GTT CGA ACT AAA CAC AGA CGT GGT CGA CCC AAA CAC
⇦ ntf021 ⇦

```
1501
CTA GGG GGG CTT TTG TCT GTG TTA TCC AAT GGA TTG CCA TTA AAT TTA TCC AAA TAA CCT
GAT CCC CCC GAA AAC AGA CAC AAT AGG TTA CCT AAC GGT AAT TTA AAT AGG TTT ATT GGA

1561
TGG ATC TGC TGG GAT TTT GCA ATA GCA TTA GCA CCT GCA CCA CCT GCA TTC AGC CCG CCT
ACC TAG ACG ACC CTA AAA CGT TAT CGT AAT CGT GGA CGT GGT GGA CGT AAG TCG GGC GGA

1621
GCT TGA TCC ATC AAA TTT TTT AGA TTT GTA GGA GTC CAG TAC AAG CCC AAT TGA AAA AAT
CGA ACT AGG TAG TTT AAA AAA TCT AAA CAT CCT CAG GTC ATG TTC GGG TTA ACT TTT TTA
                                                  Hind III
1681                                        ⇨ ✂ (Bam HI) ntf002 ⇨
GAC TCT TTT TAC TGT CAC TGT GAT TCC TTC TGC ATT TAC AAA GCT TGA AGC AAG GGA AAA
CTG AGA AAA ATG ACA GTG ACA CTA AGG AAG ACG TAA ATG TTT CGA ACT TCG TTC CCT TTT

1741                      ⇨ ✂ (Bam HI) ntf003 ⇨
TGC AAG AGT AGT AGC AAT GAG AGG TTG GAA GAA TTT TCT ATT TCT CAA AGA AGA GCT CTT
ACG TTC TCA TCA TCG TTA CTC TCC AAC CTT CTT AAA AGA TAA AGA GTT TCT TCT CGA GAA
   ⇦ ntf005 ⇦
                                                                          Sma
I
1801                                                                ⇨ ntf057 ⇨
CTC TTT TCT TTT CAA GAA ATT CTT TTC TTG AAC ATC AGA AAT TTT TAG TGA TAG CTC CCC
GAG AAA AGA AAA GTT CTT TAA GAA AAG AAC TTG TAG TCT TTA AAA ATC ACT ATC GAG GGG
                                                                ⇦ ntf056 ⇦
1861
GGG ATG CTT GAT TTG TCA CTG CTG TTC CAT TTA CGG TGG TGT TAG CAC CAT TGG TAA ATG
CCC TAC GAA CTA AAC AGT GAC GAC AAG GTA AAT GCC ACC ACA ATC GTG GTA ACC ATT TAC

1921
TAA AAG TAG CAT TCC CGT TAT TTC CGC TTA CCG TCG CAG TTT GAT TTG GAT CTG CTG GAT
ATT TTC ATC GTA AGG GCA ATA AAG GCG AAT GGC AGC GTC AAA CTA AAC CTA GAC GAC CTA

1981
TTT CTA CTG TCA CTG TGA TTC CAC TTA CCA CAG CGG GCG CTG GGG CTG CTG GTT GTG CTT
AAA GAT GAC AGT GAC ACT AAG GTG AAT GGT GTC GCC CGC GAC CCC GAC GAC CAA CAC GAA

2041
GAT CTG CTT GAT CTG CTT GAT CTG CAT TAC CTG CTT GAT CTG CAT TAC CTG CAC CTG CTG
CTA GAC GAA CTA GAC GAA CTA GAC GTA ATG GAC GAA CTA GAC GTA ATG GAC GTG GAC GAC

2101
CAT TGC TTG GCA TCC CCG TTG CAT TTG CGC TAT CGT TAT TAA ATT TTT GCA AAT AAT TTT
GTA ACG AAC CGT AGG GGC AAC GTA AAC GCG ATA GCA ATA ATT TAA AAA CGT TTA TTA AAA
                    ⇦ ntf023 ⇦
2161
CTA TCT GCT GGG ATT TTG CTT GAG CAT TAG CAT CTC CAC CAC CTG CAT TTA GCC CGC CTG
GAT AGA CGA CCC TAA AAC GAA CTC GTA ATC GTA GAG GTG GTG GAC GTA AAT CGG GCG GAC

2221
CTT GCA TCA AAT TTC CTT GAT TTG TAG GAG TCC ATT GCG TTG AAA AAT ATT TTT GGA TTT
GAA CGT AGT TTA AAG GAA CTA AAC ATC CTC AGG TAA CGC AAC TTT TTA TAA AAA CCT AAA

2281                                  ⇨ ntf019 ⇨
GTC GCT GTA ATA TTT GCA GCG GGT GTA TTT CCT GCT GCT GCA CCT GTT ACT CCT GTT CCT
CAG CGA CAT TAT AAA CGT CGC CCA CAT AAA GGA CGA CGA CGT GGA CAA TGA GGA CAA GGA

2341
  Hind III                                                      ⇨ POT009 ⇨
TCA AGC TTG GCA AAG TTA TAG AAA TTG AGG GTA TTC ACC CCT TGA TAG GCA CTC TTG ATA
AGT TCG AAC CGT TTC AAT ATC TTT AAC TCC CAT AAG TGG GGA ACT ATC CGT GAG AAC TAT

2401
TTG CCC TGG AAA ACA GTG GCA TTT GGA TTT GCT TGA CCT CCA CCT GCG CCA CCT GCT GGG
AAC GGG ACC TTT TGT CAC CGT AAA CCT AAA CGA ACT GGA GGT GGA CGC GGT GGA CGA CCC

2461
TCT GCC TTA CCA TTT ACC GTA AAA CTA TTC ACT CCG CTT GCG ATG TTT GCT ACT ATG TTC
AGA CGG AAT GGT AAA TGG CAT TTT GAT AAG TGA GGC GAA CGC TAC AAA CGA TGA TAC AAG

2521
GTC CCT GTT ACT GGA TTT GCT GGA TTT GCT GGA TTT CCA TTT ATG GTG GTG CTA GCA CCA
CAG GGA CAA TGA CCT AAA CGA CCT AAA CGA CCT AAA GGT AAA TAC CAC CAC GAT CGT GGT
```

```
2581
AAG GTA GCA CTG CCA CCA GTT CCA TCA GAT GCG GGC GCG CCG CCG CTG CCA GGG CTA GAA
TTC CAT CGT GAC GGT GGT CAA GGT AGT CTA CGC CCG CGC GGC GGC GAC GGT CCC GAT CTT
                                                          ⇦ ntf026 ⇦
2641                    ⇨ ntf018 ⇨
GTT CCA TTA TTT TGG GAT CGT GCT AAT GTT TCC TTA ATG TTT CCT GTC AAT ACT GTT TCT
CAA GGT AAT AAA ACC CTA GCA CGA TTA CAA AGG AAT TAC AAA GGA CAG TTA TGA CAA AGA
                        ⇦ ntf017 ⇦
2701
GGA TTT CCA TCT GCA CCT GCA CCT GCA CCA TCA AAT GTA ATC TTA TTT GTA GCT TGC GCA
CCT AAA GGT AGA CGT GGA CGT GGA CGT GGT AGT TTA CAT TAG AAT AAA CAT CGA ACG CGT

2761
GTT TGA TTT GGA GCT GTT GCT GGA TTT ACT TTC ACT GTG ATT GCA TTT ACC GCG CTG TGC
CAA ACT AAA CCT CGA CAA CGA CCT AAA TGA AAG TGA CAC TAA CGT AAA TGG CGC GAC ACG

2821
TTG TGC TGG ATC TGC TGG ATC TAC ACC ACC TAC TAG CTT GGC ATC CCC GTC GCA TTT GCG
AAC ACG ACC TAG ACG ACC TAG ATG TGG TGG ATG ATC GAA CCG TAG GGG CAG CGT AAA CGC
                                                    ⇦ ntf023 ⇦
2881
CTA TTG TCA TTA AAT TTT TGC AAA TAA TCG CTA ATT TGC TGG GAT TTT GCA TTA GCA TTA
GAT AAC AGT AAT TTA AAA ACG TTT ATT AGC GAT TAA ACG ACC CTA AAA CGT AAT CGT AAT

2941
GCA TTA TTT GCA CCA CCA AGC CCG CCT GCT TGA CCT ATC AAC TTT CCT AGA TTG CCA GCA
CGT AAT AAA CGT GGT GGT TCG GGC GGA CGA ACT GGA TAG TTG AAA GGA TCT AAC GGT CGT

3001
AGC CCA GTA GTA AGC CCA ATT GAA AAA TAT TTT TGG ATT TGT CGC TGT GAT ATT TGC AGC
TCG GGT CAT CAT TCG GGT TAA CTT TTT ATA AAA ACC TAA ACA GCG ACA CTA TAA ACG TCG

3061                                                    Hind III
GGG CGC TGG GGC TGC TTG ATC TGC ACC TGC TAC TGG ATT TCC TTC AAG CTT GAT TTG TGT
CCC GCG ACC CCG ACG AAC TAG ACG TGG ACG ATG ACC TAA AGG AAG TTC GAA CTA AAC ACA
                        ⇦ ntf021 ⇦
3121    Pvu II
CTG CAC CAG CTG GGT TTG TGC TAG GGG GGC TTT TGT CTG TGT TAT CCA ATG GAT TGC CAT
GAC GTG GTC GAC CCA AAC ACG ATC CCC CCG AAA ACA GAC ACA ATA GGT TAC CTA ACG GTA

3181
TAA ATT TAT CCA AAT AAC CTT GGA TCT GCT GGG ATT TTG CAA TAG CAT TAG CAC CTG CAC
ATT TAA ATA GGT TTA TTG GAA CCT AGA CGA CCC TAA AAC GTT ATC GTA ATC GTG GAC GTG

3241
CAC CTG CAT TCA GCC CGC CTG CTT GAT CCA TCA AAT TTC CTA GAT TTG TAG GAG CCC AGT
GTG GAC GTA AGT CGG GCG GAC GAA CTA GGT AGT TTA AAG GAT CTA AAC ATC CTC GGG TCA

3301
ACA AGC CCA ATT GAA AAA ATG ACT CTT TTT ACT GTC ACT GTG ATT CCT TCT GCA TTT ACA
TGT TCG GGT TAA CTT TTT TAC TGA GAA AAA TGA CAG TGA CAC TAA GGA AGA CGT AAA TGT

3361
Hind III
⇨ ✂ (Bam HI) ntf002 ⇨                              ⇨ ✂ (Bam HI) ntf003 ⇨
AAG CTT GAA GCA AGG GAA AAT GCA AGA GTA GTA GCA ATG AGA GGT TGG AAG AAT TTT CTA
TTC GAA CTT CGT TCC CTT TTA CGT TCT CAT CAT CGT TAC TCT CCA ACC TTC TTA AAA GAT
                            ⇦ ntf005 ⇦
3421
TTT CTC AAA GAA GAG CTC TTC TCT TTT CTT TTC AAG AAA TTC TTT TCT TGA ACA TCA GAA
AAA GAG TTT CTT CTC GAG AAG AGA AAA GAA AAG TTC TTT AAG AAA AGA ACT TGT AGT CTT

                        Sma I
3481                ⇨ ntf057 ⇨
ATT TTT AGT GAT AGC TCC CCG GGA TGC TTG ATT TGT CAC TGC TGT TCC ATT TAC GGT GGT
TAA AAA TCA CTA TCG AGG GGC CCT ACG AAC TAA ACA GTG ACG ACA AGG TAA ATG CCA CCA
                ⇦ ntf056 ⇦
3541
GTT AGC ACC ATT GGT AAA TGT AAA AGT AGC ATT CCC GTT ATT TCC GCT TAC CGT CGC AGT
CAA TCG TGG TAA CCA TTT ACA TTT TCA TCG TAA GGG CAA TAA AGG CGA ATG GCA GCG TCA

3601
TTG ATT TGG ATC TGC TGG ATT TTC TAC TGT CAC TGT GAT TCC ACT TAC CAC AGC GGG CGC
AAC TAA ACC TAG ACG ACC TAA AAG ATG ACA GTG ACA CTA AGG TGA ATG GTG TCG CCC GCG
```

```
3661
TGG AGC TGC TGG AGC TGC TGG TTG TGC TGG ATC TGC TTG ATT TGC TGG ATC TGC ACC TGC
ACC TCG ACG ACC TCG ACG ACC AAC ACG ACC TAG ACG AAC TAA ACG ACC TAG ACG TGG ACG

3721
TGC ATT GCT TGG CAT CCC CGT TGC ATT TGC GCT ATC GTT ATT AAA TTT TTG CAA ATA ATT
ACG TAA CGA ACC GTA GGG GCA ACG TAA ACG CGA TAG CAA TAA TTT AAA AAC GTT TAT TAA
                      ⇦ ntf023 ⇦
3781
TTC TAT CTG CTG GGA TTT TGC TTG AGC ATT AGC ATC TCC ACC ACC TGC ATT TAG CCC GCC
AAG ATA GAC GAC CCT AAA ACG AAC TCG TAA TCG TAG AGG TGG TGG ACG TAA ATC GGG CGG

3841
TGC TTG CAT CAA ATT TCC TTG ATT TGT AGG AGT CCA TTG CGT TGA AAA ATA TTT TTG GAT
ACG AAC GTA GTT TAA AGG AAC TAA ACA TCC TCA GGT AAC GCA ACT TTT TAT AAA AAC CTA

3901                                  ⇨ ntf019 ⇨
TTG TCG CTG TAA TAT TTG CAG CGG GTG TAT TTC CTG CTG CTG CAC CTG TTA CTC CTG TTC
AAC AGC GAC ATT ATA AAC GTC GCC CAC ATA AAG GAC GAC GAC GTG GAC AAT GAG GAC AAG

3961 Hind III                                                    ⇨ POT009 ⇨
CTT CAA GCT TGG CAA AGT TAT AGA AAT TGA GGG TAT TCA CCC CTT GAT AGG CAC TCT TGA
GAA GTT CGA ACC GTT TCA ATA TCT TTA ACT CCC ATA AGT GGG GAA CTA TCC GTG AGA ACT

4021
TAT TGC CCT GGA AAA CAG TGG CAT TTG GAT TTG CTT GAC CTC CAC CTG CGC CAC CTG CTG
ATA ACG GGA CCT TTT GTC ACC GTA AAC CTA AAC GAA CTG GAG GTG GAC GCG GTG GAC GAC

4081
GGT CTG CCT TAC CAT TTA CCG TAA AAC TAT TCA CTC CGC TTG CGA TGT TTG CTA CTA TGT
CCA GAC GGA ATG GTA AAT GGC ATT TTG ATA AGT GAG GCG AAC GCT ACA AAC GAT GAT ACA

4141
TCG TCC CTG TTA CTG GAT TTG CTG GAT TTG CTG GAT TTC CAT TTA TGG TGG TGC TAG CAC
AGC AGG GAC AAT GAC CTA AAC GAC CTA AAC GAC CTA AAG GTA AAT ACC ACC ACG ATC GTG

4201
CAA AGG TAG CAC TGC CAC CAG TTC CAT CAG ATG CGG GCG CGC CGC CGC TGC CAG GGC TAG
GTT TCC ATC GTG ACG GTG GTC AAG GTA GTC TAC GCC CGC GCG GCG GCG ACG GTC CCG ATC
                                                          ⇦ ntf026 ⇦
4261                          ⇨ ntf018 ⇨
AAG TTC CAT TAT TTT GGG ATC GTG CTA ATG TTT CCT TAA TGT TTC CTG TCA ATA CTG TTT
TTC AAG GTA ATA AAA CCC TAG CAC GAT TAC AAA GGA ATT ACA AAG GAC AGT TAT GAC AAA
                          ⇦ ntf017 ⇦
4321
CTG GAT TTC CAT CTG CAC CTG CAC CTG CAC CAT CAA ATG TAA TCT TAT TTG TAG CTT GCG
GAC CTA AAG GTA GAC GTG GAC GTG GAC GTG GTA GTT TAC ATT AGA ATA AAC ATC GAA CGC

4381
CAG TTT GAT TTG GAG CTG TTG CTG GAT TTA CTT TCA CTG TGA TTG CAT TTA CCG CGC TGT
GTC AAA CTA AAC CTC GAC AAC GAC CTA AAT GAA AGT GAC ACT AAC GTA AAT GGC GCG ACA

4441
GCT TGT GCT GGT GTT ACT GCT GTT CCA TCT TGC ATC AGA CTG CCA GCA CCA TTG GTA AAT
CGA ACA CGA CCA CAA TGA CGA CAA GGT AGA ACG TAG TCT GAC GGT CGT GGT AAC CAT TTA

4501
GTA AAA GTA GCA TTC CCG CCG CTT CCC TTT ACT TCA GCA GTT TGT GTA TTT TGA GCT GCT
CAT TTT CAT CGT AAG GGC GGC GAA GGG AAA TGA AGT CGT CAA ACA CAT AAA ACT CGA CGA

4561
GGA GCT GTT AGT GTC ACT GTG ATT GCT GGA GCT GCT GGA TCT GCT ACT GGA GTT CCA TCT
CCT CGA CAA TCA CAG TGA CAC TAA CGA CCT CGA CGA CCT AGA CGA TGA CCT CAA GGT AGA

4621
TTT ATC AGA CTG CCA TTG TGG GGG ATT TGC TTT GCA TAG CAT TGC CAG TAG ATC CAT CAC
AAA TAG TCT GAC GGT AAC ACC CCC TAA ACG AAA CGT ATC GTA ACG GTC ATC TAG GTA GTG

4681
TAG CAA CAC CCG CAG CGC CGC CAG TGC TAG AAG TTC CAT TAT TTT GGA ACA AAA GAT TAT
ATC GTT GTG GGC GTC GCG GCG GTC ACG ATC TTC AAG GTA ATA AAA CCT TGT TTT CTA ATA

4741
 Bam HI
TGG ATC CAG CCC ACC CTG GTG GCG TGG CAA TCA CAT TGC CCT CAA TAC TAC GGG ATC TCC
ACC TAG GTC GGG TGG GAC CAC CGC ACC GTT AGT GTA ACG GGA GTT ATG ATG CCC TAG AGG
```

```
4801
ACC TGC TGC ACC TGC ACC ATC AAA TGT AAT CTT ATT TGT TCT TAG ATT GTA GCT TGC GCA
TGG ACG ACG TGG ACG TGG TAG TTT ACA TTA GAA TAA ACA AGA ATC TAA CAT CGA ACG CGT

4861
GTT TGA TTT GCT GCT TGT GCT TGA TTT ACT GTC ACT GTG ATT GCA TTT ACC GCG CTG TGC
CAA ACT AAA CGA CGA ACA CGA ACT AAA TGA CAG TGA CAC TAA CGT AAA TGG CGC GAC ACG

4921
TTG TGC TGG ATC TGC TGG ATC TAC ACC ACC TAC TAG CTT GGC ATC CCC GTC GCA TTT GCG
AAC ACG ACC TAG ACG ACC TAG ATG TGG TGG ATG ATC GAA CCG TAG GGG CAG CGT AAA CGC
                                                          ⇦ ntf023 ⇦
4981
CTT GCA CTA TTG CCG CCA AAT TTA TCC AAA TAA CTC TGG ATC TGA TCA AAT CTT GCA TTA
GAA CGT GAT AAC GGC GGT TTA AAT AGG TTT ATT GAG ACC TAG ACT AGT TTA GAA CGT AAT

5041
GCA TTA GCA TTA TTG CCA CCA CCA AGC CCG CCT ACT TGT ATC AAA TTT CCT AGA TTT GTA
CGT AAT CGT AAT AAC GGT GGT GGT TCG GGC GGA TGA ACA TAG TTT AAA GGA TCT AAA CAT

5101
GCA AGC CCC AGC GTT GAA AAA GGA GTT GAA AAC AGT GGC TTG ACC TGC ACC ACC TGC TGC
CGT TCG GGG TCG CAA CTT TTT CCT CAA CTT TTG TCA CCG AAC TGG ACG TGG TGG ACG ACG

5161
AGC TGG TGT GAT TGT CGC ATG AGT AAG GGT TGG GAA ATT TTG CTT TAT TTG CGT TAT TCC
TCG ACC ACA CTA ACA GCG TAC TCA TTC CCA ACC CTT TAA AAC GAA ATA AAC GCA ATA AGG

5221  ⇨ ntf013 ⇨
ATG CGA TTG TCT GGA ATT GCA TTT CCA CCA CCA AGG ATG TTA AGA TTC CCA ATC AAA TTT
TAC GCT AAC AGA CCT TAA CGT AAA GGT GGT GGT TCC TAC AAT TCT AAG GGT TAG TTT AAA
      ⇦ ntf014 ⇦
5281                ⇨ ✂ (Bgl II) ntf049 ⇨
TTT GCA TTA GCA TCC CCA AGA TTC AGG GTG AAG GTT TTA GCA GTG CCA CTA CTT GCA ATC
AAA CGT AAT CGT AGG GGT TCT AAG TCC CAC TTC CAA AAT CGT CAC GGT GAT GAA CGT TAG

5341
CCA CCA AAA TCA AAA TTC AAG TTT ACT GGA TTT CCT GCA GCG CCT AGA TTT TGA TTT GCT
GGT GGT TTT AGT TTT AAG TTC AAA TGA CCT AAA GGA CGT CGC GGA TCT AAA ACT AAA CGA

5401
TGA CCT GCT TGA TCT GCA CCT GCT GGG TCT GCC TTA CCA TCC ACC GTA AAA CTA TTC ACT
ACT GGA CGA ACT AGA CGT GGA CGA CCC AGA CGG AAT GGT AGG TGG CAT TTT GAT AAG TGA
              ⇦ ntf021 ⇦
5461
CCG CTT GCG ATG TTT GCT ATT ATG TTC GTC CCT GTA GCG GTT CCT GCA CCA CCT GCT GCT
GGC GAA CGC TAC AAA CGA TAA TAC AAG CAG GGA CAT CGC CAA GGA CGT GGT GGA CGA CGA

5521
GGT TGT ACT GCT GTT CCA TCT TGC ATC AGA CTG CCA GCA CCA TTG GTA AAT GTA AAA GTA
CCA ACA TGA CGA CAA GGT AGA ACG TAG TCT GAC GGT CGT GGT AAC CAT TTA CAT TTT CAT

5581
GCA TTC CCG CCG CTT CCG CTT ACT TCA GCA TTT TGA TTT GCA CCT GCT GCT GGA GTT ACT
CGT AAG GGC GGC GAA GGC GAA TGA AGT CGT AAA ACT AAA CGT GGA CGA CGA CCT CAA TGA

5641
GTC ACT GTG ATT CCA CTT ACC GCA GCG GGC GCT GGG GGC GGG CTG GCG CGC CTA CAC CAC
CAG TGA CAC TAA GGT GAA TGG CGT CGC CCG CGA CCC CCG CCC GAC CGC GCG GAT GTG GTG

5701
CTA CAC CAC CTA CAC CAC CTA CAC CAC CTA CAC CAC CTA CTT GGC CCT GGA TGT CAT TTA
GAT GTG GTG GAT GTG GTG GAT GTG GTG GAT GTG GTG GAT GAA CCG GGA CCT ACA GTA AAT

5761       ⇨ ntf036 ⇨
AAA GCA CCT GCC TTA GTA GAA TCG TAT CTG TTA TTG TTA AGG GCC TCC CTA ACT CAT GAC
TTT CGT GGA CGG AAT CAT CTT AGC ATA GAC AAT AAC AAT TCC CGG AGG GAT TGA GTA CTG
           ⇦ ntf035 ⇦
5821
TCC GCC TGC CTA TCT GTC CTC CCT TGG TGT AGC CTC ATT GCT CGA TGA TGT CAA TGT TAT
AGG CGG ACG GAT AGA CAG GAG GGA ACC ACA TCG GAG TAA CGA GCT ACT ACA GTT ACA ATA
                                          *   Q   E   I   I   D   I   N   N

5881
TTG GCG GGG TAA AAA CAA AAA TCC CAT CAC TCA AAG TGG GAT TGA CCT TGA GAT TTT TCA
AAC CGC CCC ATT TTT GTT TTT AGG GTA GTG AGT TTC ACC CTA ACT GGA ACT CTA AAA AGT
P   P   T   F   V   F   I   G   D   S   L   T   P   N   V   K   L   N   K   L
```

```
5941
AAA TAA TCT CGA TTT TGT TTT GCA ACT CAT CAA TAA AGG TAA TTT TAT GCG GCA GAT TTT
TTT ATT AGA GCT AAA ACA AAA CGT TGA GTA GTT ATT TCC ATT AAA ATA CGC CGT CTA AAA
I   I   E   I   K   N   Q   L   E   D   I   F   T   I   K   H   P   L   N   K

6001
TGG CAA AGG AGA GCA CAT ACT GCG CAT CCC CCA CCT TCG CGT GGT ATT TGC CAT CTT GCC
ACC GTT TCC TCT CGT GTA TGA CGC GTA GGG GGT GGA AGC GCA CCA TAA ACG GTA GAA CGG
A   F   S   L   V   Y   Q   A   D   G   V   K   A   H   Y   K   G   D   Q   G

6061
    ⇨ ntf016 ⇨
CCA GAG TGG CTT TGT GCA GGA TCG CAA AAA AAT CCA CCT GCT TTG TGA GAT GTG TGA AAG
GGT CTC ACC GAA ACA CGT CCT AGC GTT TTT TTA GGT GGA CGA AAC ACT CTA CAC ACT TTC
L   T   A   K   H   L   I   A   F   F   D   V   Q   K   T   L   H   T   F   T
    ⇦ ntf015 ⇦
6121
TGG CTT GCT CAA GCA TAG GCT CAT AGA TAA TGG CCT CTT TTT TGT TGA GAT AGA TGG TTT
ACC GAA CGA GTT CGT ATC CGA GTA TCT ATT ACC GGA GAA AAA ACA ACT CTA TCT ACC AAA
A   Q   E   L   M   P   E   Y   I   I   A   E   K   K   N   L   Y   I   T   K

6181
TTT TCA GTG GGG TTT CAT AGA TCC ATT TGG CCA GAC TGG GAT TTT TGG CAT AGA GCT TGC
AAA AGT CAC CCC AAA GTA TCT AGG TAA ACC GGT CTG ACC CTA AAA ACC GTA TCT CGA ACG
K   L   P   T   E   Y   I   W   K   A   L   S   P   N   K   A   Y   L   K   G

6241
⇨ ntf060 ⇨
CTG TGT AGA TGA GCT TTT GGC TCC CTT GGT GAG TGA TCT GCA CGA AAT CCG CTT CTA TGC
GAC ACA TCT ACT CGA AAA CCG AGG GAA CCA CTC ACT AGA CGT GCT TTA GGC GAA GAT ACG
T   Y   I   L   K   Q   S   G   Q   H   T   I   Q   V   F   D   A   E   I   S

6301
TAA TAA GCC CAT TTT TCC AAA AAT CCC TCT TTA TGG CTC CTG TTG TGG TTT GCG TGG GTG
ATT ATT CGG GTA AAA AGG TTT TTA GGG AGA AAT ACC GAG GAC AAC ACC AAA CGC ACC CAC
I   L   G   N   K   W   F   D   R   K   I   A   G   T   T   T   Q   T   P   T

6361
TAG CTG CTG TGT CCT TTG CTC TCT TGC TAT CGA GTG CAA AAA CAA AAC TCC AAA AAA TCA
ATC GAC GAC ACA GGA AAC GAG AGA ACG ATA GCT CAC GTT TTT GTT TTG AGG TTT TTT AGT
A   A   T   D   K   A   R   K   S   D   L   A   F   V   F   S   W   F   I   L

6421
AAA ATG AGA AAA AAA TCT TTT GCA TGA GCG GCC TTT TTT GGA ATT TTT GAG AAA TTA CAA
TTT TAC TCT TTT TTT AGA AAA CGT ACT CGC CGG AAA AAA CCT TAA AAA CTC TTT AAT GTT
F   S   F   F   I   K   Q   M                RBS             -10?
                    orf2
6481
AAT TTG TGC GGG ATT TTA TCT GAA ATT TCC TTG CTT TTG CAT CTT TAT TGA CAT TGA TTA
TTA AAC ACG CCC TAA AAT AGA CTT TAA AGG AAC GAA AAC GTA GAA ATA ACT GTA ACT AAT

6541
⇨ ntf034 ⇨
CAG GCC CTC CGC TTG CGA TGC TTC CAT TGG TGT TTG TAT TTA GAG TTG CAT TTC CAG CGC
GTC CGG GAG GCG AAC GCT ACG AAG GTA ACC ACA AAC ATA AAT CTC AAC GTA AAG GTC GCG

6601                                                    ⇨ ✂ (Bgl II) ntf049
⇨
CAA GGA TGT TGA GAT TCC CAG TTA GTG CAT TAG CAT TAC CTG CAC CAC CAA GAT TCA GGG
GTT CCT ACA ACT CTA AGG GTC AAT CAC GTA ATC GTA ATG GAC GTG GTG GTT CTA AGT CCC
      ⇦ ntf048 ✂ (Bgl II) ⇦
6661
TGA AGG TTT TGG CTG GAT TGC CAG CAA TCC CAC CAA AAT CAA AAT TCA AGT TTA CTG GAT
ACT TCC AAA ACC GAC CTA ACG GTC GTT AGG GTG GTT TTA GTT TTA AGT TCA AAT GAC CTA

6721
TTC CTT CAG CGC CTA AGC CTG CAC CTG CTG GTT TTC CAT CCA CCA TAA AAC TAT TCA CAG
AAG GAA GTC GCG GAT TCG GAC GTG GAC GAC CAA AAG GTA GGT GGT ATT TTG ATA AGT GTC

6781
TAT TTG CAT TCC CAA GAT TCA GAG GAA TCC CCC CCT CAT CCC ATC TCC TTT ATT GTG GCT
ATA AAC GTA AGG GTT CTA AGT CTC CTT AGG GGG GGA GTA GGG TAG AGG AAA TAA CAC CGA
```

```
6841
AAA AAT CCA TAA AAA TCT ATA AAA ATC CCC ATT TGA CAA AAT TTC ATG ATA TTT TTT TGG
TTT TTA GGT ATT TTT AGA TAT TTT TAG GGG TAA ACT GTT TTA AAG TAC TAT AAA AAA ACC

6901
TAC TTG GCG GGA TTT ACC TAG GAT TTG CTG GGT TTG ATT GAG ATG TGA GTG GCA CAG CTA
ATG AAC CGC CCT AAA TGG ATC CTA AAC GAC CCA AAC TAA CTC TAC ACT CAC CGT GTC GAT

6961      ⇨ ntf010 ⇨                                          Xba I
GTT TGG TGA TGG GAG TTT GGC GGC ATC AAA AAA GCC AAA AAT TCT AGA AAA TTG TGA TAG
CAA ACC ACT ACC CTC AAA CCG CCG TAG TTT TTT CGG TTT TTA AGA TCT TTT AAC ACT ATC

7021                  Dra I
AAT CCT AGC CGG AGA TTT TAA AAT TTT CCC GCG CGC CAC TTT CAA GCG CAA GTA ATA CCA
TTA GGA TCG GCC TCT AAA ATT TTA AAA GGG CGC GCG GTG AAA GTT CGC GTT CAT TAT GGT

7081
GGG ATG AGC TTA AGG GCA AAA TCT CAT CAA AAA ATG AGC AAA AAT TAA AAT TTT CCC GCG
CCC TAC TCG AAT TCC CGT TTT AGA GTA GTT TTT TAC TCG TTT TTA ATT TTA AAA GGG CGC

7141
CTT GCG CGC TTA GTT TAT GCA AAA TTT TTC GCC TCG CTG CGC GAG ATG TAT AGA ATT TTT
GAA CGC GCG AAT CAA ATA CGT TTT AAA AAG CGG AGC GAC GCG CTC TAC ATA TCT TAA AAA
                                                    ⇦ ntf046 ≈ ⇦
7201
GGC TCA TGT GCC CAT CTT TGA GTC TTG TAA GTA CTT GAT AAT AAG CAA AGC CTT TTG CCA
CCG AGT ACA CGG GTA GAA ACT CAG AAC ATT CAT GAA CTA TTA TTC GTT TCG GAA AAC GGT

7261
GTA TTT TGG TGG GAT TGC CAG CAA TCC CCT AGG CTT TAG GAT AAG AAC CCC CCC CCC GCG
CAT AAA ACC ACC CTA ACG GTC GTT AGG GGA TCC GAA ATC CTA TTC TTG GGG GGG GGG CGC

7321        ⇨ ✂ (Bam HI) ntf031 ⇨                                  UP element
GCT CTT TGC CCT CAA ATG CCG CGC CGT GCC TGG CTG CTC CCC GCT AAA AAT TTT CCA AAA
CGA GAA ACG GGA GTT TAC GGC GCG GCA CGG ACC GAC GAG GGG CGA TTT TTA AAA GGT TTT

7381          -35                  Hind III          -10          +1
AAA TTA CAA AAT TAA TTT TTT TTT AAG CTT TTT TAT CTA AAA TAC ACT GCG TAA TTT TTT
TTT AAT GTT TTA ATT AAA AAA AAA TTC GAA AAA ATA CAT TTT ATG TCA CGC ATT AAA AAA
                                                    hsr
                                    ⇨ ntf033 ⇨                ⇨ Ⓟ ntf044 ⇨
7441                                    RBS                 M   T   K   I   S   D   V
TTT TGC GAC AAA AAA CAA AAT TAA CCT TAG GAG ATC ACA ACA TGA CAA AAA TTT CTG ATG
AAA ACG CTG TTT TTT GTT TTA ATT GGA ATC CTC TAG TGT TGT ACT GTT TTT AAA GAC TAC
                                    ⇦ ntf045 ≈ ⇦
7501
  Q   E   K   N   F   L   K   R   K   E   K   S   S   S   L   R   N   R   K   F
TTC AAG AAA AGA ATT TCT TGA AAA GAA AAG AGA AGA GCT CTT CTT TGA GAA ATA GAA AAT
AAG TTC TTT TCT TAA AGA ACT TTT CTT TTC TCT TCT CGA GAA GAA ACT CTT TAT CTT TTA

                                                                Hind III
7561                        ⇨ ntf005 ⇨
  F   Q   P   L   I   A   T   T   L   A   F   S   L   A   S   S   F   V   N   A
TCT TCC AAC CTC TCA TTG CTA CTA CTC TTG CAT TTT CCC TTG CTT CAA GCT TTG TAA ATG
AGA AGG TTG GAG AGT AAC GAT GAT GAG AAC GTA AAA GGG AAC GAA GTT CGA AAC ATT TAC
⇦ ntf003 ✂ (Bam HI) ⇦                                  ⇦ ntf002 ✂ (Bam HI)
⇦

7621 ⇨ ✂ (Bgl II) ntf028 ⇨
  A   D   A   G   N   A   G   Q   A   P   V   N   A   E   G   I   T   V   T   V
CAG CAG ATG CAG GTA ATG CAG GTC AAG CCC CAG TAA ATG CAG AAG GAA TCA CAG TGA CAG
GTC GTC TAC GTC CAT TAC GTC CAG TTC GGG GTC ATT TAC GTC TTC CTT AGT GTC ACT GTC

7681
  N   Q   A   N   K   T   A   T   V   S   G   N   N   G   N   A   T   F   T   F
TAA ATC AAG CAA ATA AAA CTG CGA CGG TAA GCG GAA ATA ACG GGA ATG CTA CTT TTA CAT
ATT TAG TTC GTT TAT TTT GAC GCT GCC ATT CGC CTT TAT TGC CCT TAC GAT GAA AAT GTA

7741
  T   N   G   A   N   T   T   V   N   G   T   A   D   P   A   V   T   A   P   N
TTA CCA ATG GTG CTA ACA CCA CCG TAA ATG GAA CAG CAG ACC CAG CAG TAA CAG CTC CAA
AAT GGT TAC CAC GAT TGT GGT GGC ATT TAC CTT GTC GTC TGG GTC GTC ATT GTC GAG GTT
```

```
7801
  I   E   V   N   I   A   N   T   V   N   N   F   T   V   D   G   K   P   A   N
ACA TTG AAG TAA ACA TCG CAA ATA CTG TGA ATA ATT TTA CGG TGG ATG GAA AAC CAG CAA
TGT AAC TTC ATT TGT AGC GTT TAT GAC ACT TAT TAA AAT GCC ACC TAC CTT TTG GTC GTT

7861
  Q   A   N   Q   N   L   G   A   E   G   K   P   V   N   L   N   F   D   F   G
ATC AAG CAA ATC AAA ATC TAG GCG CTG AAG GAA AGC CAG TAA ACT TGA ATT TTG ATT TTG
TAG TTC GTT TAG TTT TAG ATC CGC GAC TTC CTT TCG GTC ATT TGA ACT TAA AAC TAA AAC

7921
  G   I   A   S   S   G   T   A   K   T   F   T   L   N   L   G   G   A   G   N
GTG GGA TTG CAA GTA GTG GCA CTG CTA AAA CCT TCA CCC TGA ATC TTG GTG GTG CAG GTA
CAC CCT AAC GTT CAT CAC CGT GAC GAT TTT GGA AGT GGG ACT TAG AAC CAC CAC GTC CAT
                                          ⇦ ntf049 ✂ (Bgl II) ⇦      ⇦ ntf025
⇦
7981                ⇨ ✂ (Bgl II) ntf048 ⇨
  A   N   A   L   T   G   N   L   N   I   L   G   A   G   N   A   T   L   N   T
ATG CTA ATG CAC TAA CTG GGA ATC TCA ACA TCC TTG GCG CTG GAA ATG CAA CTC TAA ATA
TAC GAT TAC GTG ATT GAC CCT TAG AGT TGT AGG AAC CGC GAC CTT TAC GTT GAG ATT TAT

8041
  N   T   N   G   S   I   A   S   G   G   P   V   I   N   V   N   K   D   A   T
CAA ACA CCA ATG GAA GCA TCG CAA GCG GAG GGC CTG TAA TCA ATG TCA ATA AAG ATG CAA
GTT TGT GGT TAC CTT CGT AGC GTT CGC CTC CCG GAC ATT AGT TAC AGT TAT TTC TAC GTT
                                        ⇦ ntf034 ⇦

8101
  F   N   A   T   F   S   G   G   A   T   M   T   G   N   I   V   T   G   N   T
CAT TCA ATG CTA CTT TTA GTG GTG GTG CTA CCA TGA CTG GAA ACA TCG TAA CAG GAA ATA
GTA AGT TAC GAT GAA AAT CAC CAC CAC GAT GGT ACT GAC CTT TGT AGC ATT GTC CTT TAT

8161
  K   E   T   S   G   T   G   T   N   N   I   T   F   D   G   P   K   Q   I   P
CTA AGG AAA CCA GTG GTA CAG GAA CAA ATA ATA TTA CAT TTG ATG GTC CAA AGC AAA TCC
GAT TCC TTT GGT CAC CAT GTC CTT GTT TAT TAT AAT GTA AAC TAC CAG GTT TCG TTT AGG

8221
  H   N   G   S   L   I   K   D   G   T   A   V   T   G   Q   A   D   P   A   T
CCC ACA ATG GCA GTC TGA TAA AAG ATG GAA CAG CAG TGA CAG GTC AAG CAG ATC CAG CAA
GGG TGT TAC CGT CAG ACT ATT TTC TAC CTT GTC GTC ACT GTC CAG TTC GTC TAG GTC GTT

8281
  V   L   T   G   N   I   S   T   Y   G   G   I   N   N   V   T   F   E   K   G
CAG TAT TGA CAG GAA ACA TTA GCA CGT ACG GCG GTA TCA ACA ATG TAA CTT TTG AGA AAG
GTC ATA ACT GTC CTT TGT AAT CGT GCA TGC CGC CAT AGT TGT TAC ATT GAA AAC TCT TTC

8341
  T   M   K   G   D   I   I   A   G   N   A   T   G   Q   S   L   G   M   N   V
GTA CGA TGA AAG GGG ATA TCA TAG CAG GTA ATG CGA CGG GGC AAT CTC TGG GAA TGA ATG
CAT GCT ACT TTC CCC TAT AGT ATC GTC CAT TAC GCT GCC CCG TTA GAG ACC CTT ACT TAC

8401      Dra I
  V   T   F   K   E   Q   G   V   H   Y   T   G   N   V   I   A   S   G   T   G
TGG TAA CCT TTA AAG AAC AAG GCG TTC ATT ACA CAG GCA ACG TAA TCG CTT CAG GAA CTG
ACC ATT GGA AAT TTC TTG TTC CGC AAG TAA TGT GTC CGT TGC ATT AGC GAA GTC CTT GAC

8461
  G   V   N   N   T   L   N   F   G   N   A   T   V   D   A   T   N   G   G   N
GTG GAG TGA ATA ACA CCC TGA ATT TTG GGA ATG CTA CTG TGG ATG CCA CCA ATG GAG GAA
CAC CTC ACT TAT TGT GGG ACT TAA AAC CCT TAC GAT GAC ACC TAC GGT GGT TAC CTC CTT

8521                Eco RI
  T   L   I   I   Q   N   S   G   I   T   F   N   N   T   N   G   V   N   N   S
ACA CTC TAA TCA TCC AGA ATT CTG GAA TCA CAT TCA ATA ACA CCA ATG GAG TGA ATA ATT
TGT GAG ATT AGT AGG TCT TAA GAC CTT AGT GTA AGT TAT TGT GGT TAC CTC ACT TAT TAA

8581                                  Pvu II
  P   T   L   T   H   A   T   I   T   P   A   A   A   G   G   D   P   A   N   Q
CTC CAA CCC TTA CTC ATG CGA CAA TCA CAC CAG CTG CAG CAG GTG GAG ATC CAG CAA ATC
GAG GTT GGG AAT GAG TAC GCT GTT AGT GTG GTC GAC GTC GTC CAC CTC TAG GTC GTT TAG

8641
  A   T   V   F   Q   G   N   I   K   S   A   Y   Q   G   V   N   T   L   N   F
AGG CCA CTG TTT TCC AGG GCA ATA TCA AGA GTG CCT ATC AAG GGG TGA ATA CCC TCA ATT
TCC GGT GAC AAA AGG TCC CGT TAT AGT TCT CAC GGA TAG TTC CCC ACT TAT GGG AGT TAA
                                    ⇦ POT009 ⇦
```

```
8701                 Hind III
  Y   N   F   A   K   L   E   G   T   P   A   N   K   A   N   P   A   P   A   A
TCT ATA ACT TTG CCA AGC TTG AAG GAA CTC CAG CAA ATA AAG CAA ATC CAG CGC CCG CTG
AGA TAT TGA AAC GGT TCG AAC TTC CTT GAG GTC GTT TAT TTC GTT TAG GTC GCG GGC GAC

8761
  N   I   T   A   T   N   N   G   A   N   N   I   V   F   T   D   G   G   L   V
CAA ATA TCA CAG CGA CAA ATA ATG GGG CAA ATA ACA TCG TAT TCA CAG ATG GAG GTT TGG
GTT TAT AGT GTC GCT GTT TAT TAC CCC GTT TAT TGT AGC ATA AGT GTC TAC CTC CAA ACC

8821
  N   A   N   L   T   S   T   L   D   Q   G   I   N   T   L   V   M   N   T   N
TGA ATG CCA ATC TCA CGT CTA CAC TGG ACC AAG GAA TCA ACA CAC TTG TGA TGA ATA CAA
ACT TAC GGT TAG AGT GCA GAT GTG ACC TGG TTC CTT AGT TGT GTG AAC ACT ACT TAT GTT

8881
  N   I   V   T   N   P   I   L   L   T   G   N   V   V   T   N   T   P   G   W
ACA ACA TCG TAA CTA ACC CCA TTT TGC TTA CTG GAA ACG TAG TAA CAA ACA CAC CAG GGT
TGT TGT AGC ATT GAT TGG GGT AAA ACG AAT GAC CTT TGC ATC ATT GTT TGT GTG GTC CCA
                                  ⇦ ntf050 ⇦
8941                                                ⇨ ntf026 ⇨
  A   G   S   N   T   L   L   F   Q   N   N   G   T   S   S   T   G   G   N   A
GGG CTG GAT CCA ATA CTC TTT TGT TCC AAA ATA ATG GAA CTT CTA GCA CTG GCG GGA ATG
CCC GAC CTA GGT TAT GAG AAA ACA AGG TTT TAT TAC CTT GAA GAT CGT GAC CGC CCT TAC

9001
  M   Q   T   L   T   N   Q   V   A   Y   V   G   N   I   V   A   N   G   G   S
CTA TGC AAA CGC TAA CAA ATC AAG TGG CCT ATG TGG GAA ATA TCG TAG CCA ATG GAG GGT
GAT ACG TTT GCG ATT GTT TAG TTC ACC GGA TAC ACC CTT TAT AGC ATC GGT TAC CTC CCA
                                      ⇦ POT008 ⇦
9061
  V   Q   A   I   F   S   N   T   Y   W   A   P   T   N   L   K   D   L   K   E
CTG TAC AAG CAA TCT TTA GTA ATA CTT ACT GGG CTC CTA CAA ATC TAA AAG ATT TGA AGG
GAC ATG TTC GTT AGA AAT CAT TAT GAA TGA CCC GAG GAT GTT TAG ATT TTC TAA ACT TCC

9121
  Q   A   G   G   L   N   A   A   G   A   A   G   A   N   A   R   A   N   A   Q
AAC AAG CAG GCG GGC TAA ATG CAG CAG GTG CAG CAG GTG CTA ATG CTA GGG CTA ATG CTC
TTG TTC GTC CGC CCG ATT TAC GTC GTC CAC GTC GTC CAC GAT TAC GAT CCC GAT TAC GAG

9181
  A   K   S   Q   Q   I   Q   G   Y   L   D   K   F   N   G   N   S   A   N   A
AAG CAA AAT CCC AGC AGA TCC AAG GTT ATT TGG ATA AAT TTA ATG GCA ATA GCG CAA ATG
TTC GTT TTA GGG TCG TCT AGG TTC CAA TAA ACC TAT TTA AAT TAC CGT TAT CGC GTT TAC

9241
  T   G   N   L   T   A   T   N   G   G   T   A   T   L   V   L   R   N   T   T
CGA CGG GGA ACT TAA CTG CCA CAA ATG GAG GTA CTG CTA CTT TGG TGC TAC GAA ACA CCA
GCT GCC CCT TGA ATT GAC GGT GTT TAC CTC CAT GAC GAT GAA ACC ACG ATG CTT TGT GGT

9301
  T   L   A   N   L   P   R   Q   A   A   Q   Y   N   V   T   V   G   G   N   N
CTA CTC TTG CAA ATC TAC CAC GTC AAG CAG CGC AAT ACA ACG TCA CTG TAG GTG GAA ATA
GAT GAG AAC GTT TAG ATG GTG CAG TTC GTC GCG TTA TGT TGC AGT GAC ATC CAC CTT TAT

9361
  S   S   A   N   I   V   L   E   A   P   V   N   A   S   A   T   I   T   Y   G
ACA GTA GCG CAA ATA TCG TCT TAG AGG CAC CTG TAA ATG CAA GTG CAA CCA TCA CTT ATG
TGT CAT CGC GTT TAT AGC AGA ATC TCC GTG GAC ATT TAC GTT CAC GTT GGT AGT GAA TAC

9421
  G   Y   Y   L   G   G   N   G   T   S   N   Y   V   W   N   G   S   Q   N   T
GAG GTT ATT ATT TAG GTG GAA ATG GGA CTA GTA ATT ATG TTT GGA ATG GTA GTC AAA ATA
CTC CAA TAA TAA ATC CAC CTT TAC CCT GAT CAT TAA TAC AAA CCT TAC CAT CAG TTT TAT

9481
  S   S   V   N   L   I   F   A   N   A   D   N   R   G   T   P   T   L   N   G
CTT CTA GTG TGA ATT TGA TCT TTG CAA ACG CTG ATA ATC GCG GGA CTC CTA CGC TTA ATG
GAA GAT CAC ACT TAA ACT AGA AAC GTT TGC GAC TAT TAG CGC CCT GAG GAT GCG AAT TAC

9541
  A   T   G   S   S   T   L   V   S   D   A   F   G   G   Q   F   R   N   D   L
GGG CAA CAG GGA GTA GTA CGC TAG TAA GCG ATG CGT TTG GAG GGC AGT TTA GAA ATG ATC
CCC GTT GTC CCT CAT CAT GCG ATC ATT CGC TAC GCA AAC CTC CCG TCA AAT CTT TAC TAG
```

```
9601                                                          Bam HI
  G   A   G   K   V   L   G   V   T   Y   Q   N   G   I   Q   M   S   L   S   D
TAG GCG CTG GGA AGG TGC TAG GAG TGA CTT ATC AAA ACG GGA TTC AAA TGA GCC TAA GTG
ATC CGC GAC CCT TCC ACG ATC CTC ACT GAA TAG TTT TGC CCT AAG TTT ACT CGG ATT CAC

9661
  K   N   V   T   L   Q   G   Q   N   G   L   Y   S   G   S   F   M   A   F   F
ATA AAA ATG TGA CCT TGC AGG GAC AAA ATG GGC TTT ATT CGG GGT CTT TCA TGG CGT TTT
TAT TTT TAC ACT GGA ACG TCC CTG TTT TAC CCG AAA TAA GCC CCA GAA AGT ACC GCA AAA

9721
  K   D   A   I   L   A   K   I   A   K   V   D   S   N   A   E   F   A   T   Q
TTA AGG ATG CTA TCT TAG CAA AAA TTG CGA AGG TGG ACT CCA ATG CAG AAT TTG CGA CTC
AAT TCC TAC GAT AGA ATC GTT TTT AAC GCT TCC ACC TGA GGT TAC GTC TTA AAC GCT GAG

9781                                                          ⇨ ntf026 ⇨
  G   I   P   L   N   V   S   L   V   K   S   G   N   G   T   S   S   P   G   S
AAG GAA TCC CTC TAA ATG TTA GCC TCG TGA AGA GTG GTA ATG GAA CTT CTA GCC CTG GCA
TTC CTT AGG GAG ATT TAC AAT CGG AGC ACT TCT CAC CAT TAC CTT GAA GAT CGG GAC CGT

9841
  G   G   N   S   F   V   N   N   I   T   L   E   G   V   A   V   G   S   I   T
GCG GCG GGA ATA GCT TTG TGA ATA ACA TTA CTT TGG AAG GGG TTG CAG TAG GAA GCA TCA
CGC CGC CCT TAT CGA AAC ACT TAT TGT AAT GAA ACC TTC CCC AAC GTC ATC CTT CGT AGT

9901
  A   L   T   N   K   Q   A   T   G   T   N   G   M   N   N   T   S   G   I   V
CTG CTC TTA CAA ACA AGC AAG CTA CAG GGA CGA ACG GCA TGA ATA ACA CCA GTG GGA TTG
GAC GAG AAT GTT TGT TCG TTC GAT GTC CCT GCT TGC CGT ACT TAT TGT GGT CAC CCT AAC

9961
  N   L   V   L   K   S   D   S   V   L   L   G   T   I   A   G   E   N   Q   K
TGA ATC TCG TCT TAA AAA GTG ATA GTG TTT TGC TTG GCA CCA TTG CGG GCG AAA ATC AAA
ACT TAG AGC AGA ATT TTT CAC TAT CAC AAA ACG AAC CGT GGT AAC GCC CGC TTT TAG TTT

10021
  G   L   T   M   N   M   Q   L   N   Q   G   A   K   L   I   L   Q   N   S   G
AAG GCC TCA CAA TGA ATA TGC AAC TCA ATC AAG GGG CTA AGC TGA TTT TGC AAA ATA GCG
TTC CGG AGT GTT ACT TAT ACG TTG AGT TAG TTC CCC GAT TCG ACT AAA ACG TTT TAT CGC

10081
  A   G   T   G   G   D   V   A   L   N   N   L   T   I   A   S   G   N   N   G
GCG CTG GGA CAG GGG GAG ATG TGG CGC TCA ATA ACT TGA CAA TCG CTT CGG GAA ATA ACG
CGC GAC CCT GTC CCC CTC TAC ACC GCG AGT TAT TGA ACT GTT AGC GAA GCC CTT TAT TGC

10141
  N   G   N   N   G   A   A   V   T   F   Q   G   G   S   V   S   F   D   K   N
GGA ATG GAA ATA ACG GGG CAG CAG TGA CAT TCC AGG GTG GAA GCG TGA GCT TTG ATA AAA
CCT TAC CTT TAT TGC CCC GTC GTC ACT GTA AGG TCC CAC CTT CGC ACT CGA AAC TAT TTT

10201
  Q   A   N   D   Y   T   A   L   Q   N   N   T   V   I   D   L   A   T   G   G
ATC AAG CAA ATG ACT ATA CTG CAC TCC AGA ATA ATA CCG TCA TCG ACC TAG CCA CAG GTG
TAG TTC GTT TAC TGA TAT GAC GTG AGG TCT TAT TAT GGC AGT AGC TGG ATC GGT GTC CAC

10261
  G   S   N   N   V   P   S   R   T   W   F   N   L   L   T   V   G   Q   A   N
GGG GTA GCA ATA ATG TCC CAA GTA GAA CAT GGT TCA ACC TCC TTA CTG TCG GTC AAG CCA
CCC CAT CGT TAT TAC AGG GTT CAT CTT GTA CCA AGT TGG AGG AAT GAC AGC CAG TTC GGT

10321                                                    Hind III
  S   S   N   T   T   T   A   S   D   G   Q   Q   A   S   G   L   G   G   N   N
ACT CTA GTA ACA CGA CTA CGG CTA GTG ATG GGC AAC AAG CTT CAG GAC TTG GTG GGA ACA
TGA GAT CAT TGT GCT GAT GCC GAT CAC TAC CCG TTG TTC GAA GTC CTG AAC CAC CCT TGT

10381     Dra I
  A   L   F   K   V   Y   V   N   A   D   A   N   Q   G   N   G   A   G   G   G
ATG CGC TCT TTA AAG TCT ATG TAA ACG CAG ATG CTA ACC AAG GAA ATG GTG CTG GGG GTG
TAC GCG AGA AAT TTC AGA TAC ATT TGC GTC TAC GAT TGG TTC CTT TAC CAC GAC CCC CAC

10441                                  Eco RI
  R   G   N   A   T   L   N   G   Q   N   S   F   N   G   S   G   L   Y   G   N
GGC GGG GGA ACG CGA CTC TAA ATG GTC AGA ATT CTT TCA ACG GAT CTG GAC TCT ATG GAA
CCG CCC CCT TGC GCT GAG ATT TAC CAG TCT TAA GAA AGT TGC CTA GAC CTG AGA TAC CTT
```

```
10501
  I   Y   S   D   R   V   I   V   Y   Q   T   Q   E   Q   H   F   C   D   R   I
ACA TCT ATA GTG ACC GTG TGA TTG TGT ACC AAA CCC AGG AAC AAC ACT TCT GTG ACA GAA
TGT AGA TAT CAC TGG CAC ACT AAC ACA TGG TTT GGG TCC TTG TTG TGA AGA CAC TGT CTT

10561
  S   P   N   P   R   Q   W   K   S   Y   G   V   R   Y   H   G   G   G   T   E
TAT CTC CAA ATC CTA GGC AAT GGA AAT CTT ATG GTG TGA GAT ACC ATG GCG GAG GTA CTG
ATA GAG GTT TAG GAT CCG TTA CCT TTA GAA TAC CAC ACT CTA TGG TAC CGC CTC CAT GAC

10621
  R   A   G   N   V   A   V   A   T   V   K   N   E   G   G   Q   A   S   V   N
AGA GGG CAG GTA ACG TTG CGG TAG CCA CTG TAA AAA ATG AAG GTG GGC AGG CGA GTG TGA
TCT CCC GTC CAT TGC AAC GCC ATC GGT GAC ATT TTT TAC TTC CAC CCG TCC GCT CAC ACT

10681
  F   T   T   V   G   S   V   I   G   F   D   V   F   D   A   K   L   T   A   V
ATT TCA CAA CTG TGG GTT CTG TCA TTG GTT TTG ATG TGT TTG ATG CAA AAC TTA CTG CGG
TAA AGT GTT GAC ACC CAA GAC AGT AAC CAA AAC TAC ACA AAC TAC GTT TTG AAT GAC GCC

10741
  K   T   N   A   Y   G   K   V   E   T   N   N   A   N   N   A   G   N   S   T
TAA AAA CAA ATG CCT ATG GTA AGG TAG AAA CAA ATA ATG CTA ATA ACG CAG GAA ATA GTA
ATT TTT GTT TAC GGA TAC CAT TCC ATC TTT GTT TAT TAC GAT TAT TGC GTC CTT TAT CAT

10801
  P   A   P   G   L   G   S   I   P   G   L   G   G   T   G   G   T   S   S   G
CTC CAG CGC CAG GGC TAG GTA GTA TCC CAG GAC TAG GTG GGA CTG GTG GCA CAA GCA GTG
GAG GTC GCG GTC CCG ATC CAT CAT AGG GTC CTG ATC CAC CCT GAC CAC CGT GTT CGT CAC

10861
  N   G   T   G   G   S   Q   D   Q   A   N   A   Q   D   Y   T   T   Y   F   I
GCA ATG GAA CTG GTG GTA GCC AAG ATC AGG CAA ATG CTC AAG ACT ATA CTA CTT ACT TCA
CGT TAC CTT GAC CAC CAT CGG TTC TAG TCC GTT TAC GAG TTC TGA TAT GAT GAA TGA AGT

10921
  S   Q   A   V   A   N   T   S   E   A   N   Q   L   A   T   A   T   A   L   A
TCA GTC AAG CAG TGG CAA ATA CTT CAG AGG CCA ACC AAC TCG CAA CAG CAA CAG CAC TTG
AGT CAG TTC GTC ACC GTT TAT GAA GTC TCC GGT TGG TTG AGC GTT GTC GTT GTC GTG AAC

                                        Cla I
10981                                    ⇨ ✂ (Pvu I/Bgl II) ntf008/ntf011 ⇨
  S   N   Y   Y   L   Y   L   A   N   I   D   S   L   N   K   R   M   G   E   L
CTA GCA ACT ACT ATC TCT ACC TAG CAA ACA TCG ATA GTC TCA ATA AGC GTA TGG GTG AGC
GAT CGT TGA TGA TAG AGA TGG ATC GTT TGT AGC TAT CAG AGT TAT TCG CAT ACC CAC TCG
            ⇦ ntf009/ntf012 ✂ (Pvu I/Bgl II) ⇦
11041
  R   S   N   P   R   S   N   G   F   W   M   R   M   F   N   G   M   Q   T   T
TTC GCA GCA ATC CGC GCA GTA ATG GCT TCT GGA TGC GTA TGT TTA ATG GTA TGC AAA CCA
AAG CGT CGT TAG GCG CGT CAT TAC CGA AGA CCT ACG CAT ACA AAT TAC CAT ACG TTT GGT

11101
  K   F   A   L   Q   T   T   S   I   Y   T   T   V   Q   A   G   W   D   H   V
CAA AAT TTG CAC TTC AGA CTA CCT CAA TCT ACA CCA CAG TAC AGG CAG GAT GGG ATC ATG
GTT TTA AAC GTG AAG TCT GAT GGA GTT AGA TGT GGT GTC ATG TCC GTC CTA CCC TAG TAC

11161
  F   G   S   E   G   G   N   D   F   L   G   F   A   V   A   Y   A   G   A   A
TAT TTG GCA GCG AGG GTG GAA ATG ACT TTT TAG GTT TTG CTG TGG CTT ATG CAG GTG CAG
ATA AAC CGT CGC TCC CAC CTT TAC TGA AAA ATC CAA AAC GAC ACC GAA TAC GTC CAC GTC

11221
  M   S   S   E   K   K   E   Q   L   V   N   G   A   Q   K   G   V   K   S   S
CGA TGA GCT CTG AGA AGA AAG AAC AGC TAG TAA ATG GTG CAC AAA AGG GAG TAA AAT CCA
GCT ACT CGA GAC TCT TCT TTC TTG TCG ATC ATT TAC CAC GTG TTT TCC CTC ATT TTA GGT

11281
  G   G   N   A   F   E   I   S   L   Y   N   S   Y   V   Q   D   G   A   A   S
GCG GTG GAA ATG CCT TTG AAA TCT CGC TCT ACA ACT CCT ATG TAC AAG ATG GTG CTG CTT
CGC CAC CTT TAC GGA AAC TTT AGA GCG AGA TGT TGA GGA TAC ATG TTC TAC CAC GAC GAA

11341
  S   T   D   F   K   Y   G   F   Y   S   D   S   V   A   K   F   S   F   L   W
CTA GCA CAG ATT TCA AGT ATG GTT TTT ATA GTG ATA GCG TGG CAA AAT TCA GCT TCT TGT
GAT CGT GTC TAA AGT TCA TAC CAA AAA TAT CAC TAT CGC ACC GTT TTA AGT CGA AGA ACA
```

```
11401
      Hind III
  N   K   L   T   M   F   G   E   D   S   S   P   N   M   Q   N   F   G   F   T
GGA ACA AGC TTA CAA TGT TTG GTG AGG ACA GCT CTC CTA ACA TGC AAA ACT TTG GTT TCA
CCT TGT TCG AAT GTT ACA AAC CAC TCC TGT CGA GAG GAT TGT ACG TTT TGA AAC CAA AGT

11461
  F   S   Q   E   I   G   Y   R   F   L   L   G   N   H   N   E   W   Y   I   T
CCT TCT CTC AAG AGA TTG GTT ATC GCT TCT TGC TAG GAA ATC ACA ACG AGT GGT ATA TCA
GGA AGA GAG TTC TCT AAC CAA TAG CGA AGA ACG ATC CTT TAG TGT TGC TCA CCA TAT AGT

11521
  P   Q   G   Q   V   A   L   G   Y   F   N   Q   S   N   I   K   Q   T   L   G
CTC CAC AAG GGC AAG TTG CTT TAG GTT ATT TCA ACC AAA GCA ATA TCA AGC AAA CCC TAG
GAG GTG TTC CCG TTC AAC GAA ATC CAA TAA AGT TGG TTT CGT TAT AGT TCG TTT GGG ATC

11581
  S   H   W   L   K   G   E   Q   S   S   I   F   T   V   Q   G   R   I   G   S
GAA GCC ACT GGC TAA AAG GCG AGC AAA GTT CTA TCT TCA CAG TGC AGG GGC GAA TTG GAA
CTT CGG TGA CCG ATT TTC CGC TCG TTT CAA GAT AGA AGT GTC ACG TCC CCG CTT AAC CTT

11641
  N   F   G   Y   R   F   N   Q   F   T   E   D   K   G   W   A   S   E   L   Y
GCA ACT TTG GTT ATA GAT TTA ATC AAT TCA CTG AAG ACA AGG GCT GGG CTT CAG AGC TTT
CGT TGA AAC CAA TAT CTA AAT TAG TTA AGT GAC TTC TGT TCC CGA CCC GAA GTC TCG AAA

11701
  L   G   L   W   Y   I   G   D   Y   I   S   G   G   N   L   T   L   V   S   D
ATT TGG GCT TGT GGT ACA TCG GCG ATT ATA TCA GTG GTG GCA ATC TTA CCC TCG TGT CTG
TAA ACC CGA ACA CCA TGT AGC CGC TAA TAT AGT CAC CAC CGT TAG AAT GGG AGC ACA GAC

11761
  L   G   S   V   N   T   L   R   T   L   S   S   T   G   R   F   A   F   N   I
ACC TAG GTT CTG TAA ACA CTT TAA GGA CTT TGA GCT CTA CTG GTA GAT TTG CCT TTA ACA
TGG ATC CAA GAC ATT TGT GAA ATT CCT GAA ACT CGA GAT GAC CAT CTA AAC GGA AAT TGT

11821
  G   T   N   F   V   V   K   D   N   H   R   F   Y   F   D   F   E   R   S   F
TTG GTA CAA ACT TCG TCG TCA AAG ATA ATC ATA GAT TCT ACT TTG ATT TTG AAA GAA GCT
AAC CAT GTT TGA AGC AGC AGT TTC TAT TAG TAT CTA AGA TGA AAC TAA AAC TTT CTT CGA

11881
  G   G   K   I   I   T   D   Y   Q   F   N   I   G   Y   R   Y   N   F   G   E
TTG GAG GCA AAA TCA TCA CAG ATT ACC AAT TCA ACA TTG GCT ATC GCT ATA ACT TTG GCG
AAC CTC CGT TTT AGT AGT GTC TAA TGG TTA AGT TGT AAC CGA TAG CGA TAT TGA AAC CGC

11941
  N   R   K   Y   V   S   L   L   A   G   S   M   K   D   T   I   K   K  -D   D
AAA ACA GAA AAT ACG TTT CTC TTC TTG CAG GTA GTA TGA AAG ACA CTA TCA AAA AAG ATG
TTT TGT CTT TTA TGC AAA GAG AAG AAC GTC CAT CAT ACT TTC TGT GAT AGT TTT TTC TAC

12001
  K   K   E   N   K   E   E   T   E   E   I   *
ATA AGA AAG AAA ACA AAG AAG AGA CAG AAG AAA TTG AGT GAT TTT CCA AAA AAT TAC TCA
TAT TCT TTC TTT TGT TTC TTC TCT GTC TTC TTT AAC TCA CTA AAA GGT TTT TTA ATG AGT

12061
TAG TGA TTC AGT CCG CCT TGG GCT GAA GTA AGC ACG AGG GGA ATA TGG AGG GGA TTC CGG
ATC ACT AAG TCA GGC GGA ACC CGA CTT CAT TCG TGC TCC CCT TAT ACC TCC CCT AAG GCC
                          ⇦ ntf037 ✂ (Bgl II) ⇦
12121              Eco RI Xba I
GGG CTT TTG TCT TGG AAT TCT AGA AAT TTT TTG GTG CGC AAC ACT GCC TCT AAC AAA AGC
CCC GAA AAC AGA ACC TTA AGA TCT TTA AAA AAC CAC GCG TTG TGA CGG AGA TTG TTT TCG

12181                                        Xba I
TCT GGT TCA AAA GGC TTT TGA TGG TGG CGT GAA TCT AGA CAA GCC TAG TAG CAA TGC AGT
AGA CCA AGT TTT CCG AAA ACT ACC ACC GCA CTT AGA TCT GTT CGG ATC ATC GTT ACG TCA

12241
CAG TGT TTG TAT GAC TAG TCT CAT GTA GAA GGT GTG GAT TAG GTT CTG GGC GTG AGT GAA
GTC ACA AAC ATA CTG ATC AGA GTA CAT CTT CCA CAC CTA ATC CAA GAC CCG CAC TCA CTT

12301
GAG GAT TTT AAG ACA TGC TTC CTC CTT GCT TTG TGT GAT ATG AGG AAG GGC CTC TGG GCT
CTC CTA AAA TTC TGT ACG AAG GAG GAA CGA AAC ACA CTA TAC TCC TTC CCG GAG ACC CGA
```

```
12361        Dra I
ACT TAT AGG TTT AAA TTT TGT AAT TTC TCA AAA ACA CAG CTT TTA GAA GTG ATA GAA GTG
TGA ATA TCC AAA TTT AAA ACA TTA AAG AGT TTT TGT GTC GAA AAT CTT CAC TAT CTT CAC

12421
GTG ATT TTG AAA GGC TTT GAG GAT TTT TAT CCT TGA GGC TCC TTT ACT AGT CTC TGG CTT
CAC TAA AAC TTT CCG AAA CTC CTA AAA ATA GGA ACT CCG AGG AAA TGA TCA GAG ACC GAA

12481                                                    ⇨ ntf051 ⇨
AAG ATG TGA GCT TGT CAA AGC GGG GGT AAA AAC TAA TGC ACC TGT TTG AGT CTA GGT CAT
TTC TAC ACT CGA ACA GTT TCG CCC CCA TTT TTG ATT ACG TGG ACA AAC TCA GAT CCA GTA

12541
GAC TTA GGT TTG AGA CTC TCT TGC AAT AGC GGG ATT GAG CTA GCT CAA TCC CTA AGT TGA
CTG AAT CCA AAC TCT GAG AGA ACG TTA TCG CCC TAA CTC GAT CGA GTT AGG GAT TCA ACT

12601
CAA TTT GTC GCA AAA GAT GCT CCA TAA AAA GGA GCA TCA CTA TTT TAC TAG GGT TAA ATA
GTT AAA CAG CGT TTT CTA CGA GGT ATT TTT CCT CGT AGT GAT AAA ATG ATC CCA ATT TAT

12661
AAC TGC CAT AAG TAG CAG GAA TTT TTG CTA GGG CTA ATA ATC TAC CAA GAT GCA ACC ACA
TTG ACG GTA TTC ATC GTC CTT AAA AAC GAT CCC GAT TAT TAG ATG GTT CTA CGT TGG TGT
12721
CAA CTC TGC AAT TTT GCG CCA AGA GCG ATG AAA AAG GCC TCA AAA ATA CAG CAC AAA ACT
GTT GAG ACG TTA AAA CGC GGT TCT CGC TAC TTT TTC CGG AGT TTT TAT GTC GTG TTT TGA
              ⇦ ntf055 ⇦
12781
GGC TCT CTA GCG CAA AAA ATT TCA CGA TTT TGT AAT CCC ACC AAA ATC AAA AAC ACA GCA
CCG AGA GAT CGC GTT TTT TAA AGT GCT AAA ACA TTA GGG TGG TTT TAG TTT TTG TGT CGT

12841
AAA CCT ACA AAG ACA TAA AAG ATG CAC AAA ACC ACA GGG CAA AAT ATT GAA GTC AAT ATT
TTT GGA TGT TTC TGT ATT TTC TAC GTG TTT TGG TGT CCC GTT TTA TAA CTT CAG TTA TAA

12901
GCA AAT ACT GTG AAT AAT TTT ACG GTA AAT GGT AAG GCA GAT CCA AGG TTA TTT GGA TAA
CGT TTA TGA CAC TTA TTA AAA TGC CAT TTA CCA TTC CGT CTA GGT TCC AAT AAA CCT ATT

12961                                                              Pvu II
ATT TAA TGG CAA TCC ATT GGA TAA CAC AGA CAA AAG CCC CCC TAG CAC AAA CCC AGC TGG
TAA ATT ACC GTT AGG TAA CCT ATT GTG TCT GTT TTC GGG GGG ATC GTG TTT GGG TCG ACC

13021             Hind III
TGC AGA CAC AAA TCA AGC TTG TAG GAA CAG GAG TAA CAG GTG CTG CAA ATA TTA TTT GCT
ACG TCT GTG TTT AGT TCG AAC ATC CTT GTC CTC ATT GTC CAC GAC GTT TAT AAT AAA CGA

13081                                                  ⇨ ✂ (Bam HI) ntf002
⇨
CTT ATT TGT AGC TTG TAC TGG ATC TAC TTT CAC TGT GAT TGT ATC TGC TGC ATT TAC AAA
GAA TAA ACA TCG AAC ATG ACC TAG ATG AAA GTG ACA CTA ACA TAG ACG ACG TAA ATG TTT

13141                                                ⇨ ✂ (Bam HI) ntf003 ⇨
GCT TGA AGC AAG GGA AAA TGC AAG AGT AGT AGC AAT GAG AGG TTG GAA GAA TTT TCT ATT
CGA ACT TCG TTC CCT TTT ACG TTC TCA TCA TCG TTA CTC TCC AAC CTT CTT AAA AGA TAA
                        ⇦ ntf005 ⇦
13201
TCT CAA AGA AGA GCT CTT CTC TTT TCT TTT CAA GAA ATT CTT TTC TTG ACA TCA GAA ATT
AGA GTT TCT TCT CGA GAA GAG AAA AGA AAA GTT CTT TAA GAA AAG AAC TGT AGT CTT TAA
                   Sma I
13261       ⇨ ntf057 ⇨
TTT AGT GAT AGC TCC CCG GGA TGC TTG ATT TGT CAC TGC GAT TCC ATC TCG CAT TAG TTT
AAA TCA CTA TCG AGG GGC CCT ACG AAC TAA ACA GTG ACG CTA AGG TAG AGC GTA ATC AAA
            ⇦ ntf056 ⇦
13321
TTA GTG GCC ACT TGA TTT GCT TTG CAT AGC ATT GCC AGT AGA TCC ATC ACT AGC AAC ACC
AAT CAC CGG TGA ACT AAA CGA AAC GTA TCG TAA CGG TCA TCT AGG TAG TGA TCG TTG TGG

13381
CGC AGC GCC GCC AGT GCT AGA AGT TCC ATT ATT TTG GAG GAT TTG TCT AGT GTT GCG ATT
GCG TCG CGG CGG TCA CGA TCT TCA AGG TAA TAA AAC CTC CTA AAC AGA TCA CAA CGC TAA

13441
TCA TTC CCC CAC CTC TAC TGT TGC TGG ATC TGC TTG ACC TGC TAC TGG AGT TCC ATC TTT
AGT AAG GGG GTG GAG ATG ACA ACG ACC TAG ACG AAC TGG ACG ATG ACC TCA AGG TAG AAA
```

```
13501
TAT CAG ACT GCC ATT GTG GGG GAT TTG CTT TGC ATA GCA TTG CCA GTA GAT CCA TCA CTA
ATA GTC TGA CGG TAA CAC CCC CTA AAC GAA ACG TAT CGT AAC GGT CAT CTA GGT AGT GAT

13561
GCA ACA CCC GCA GCG CCG CCG CTG TTA GTG CTA GAA GTT CCA TTA TTT TGG ATT ATT GCC
CGT TGT GGG CGT CGC GGC GGC GAC AAT CAC GAT CTT CAA GGT AAT AAA ACC TAA TAA CGG

13621                                                  Bam HI
TAG TGT TGC GAT TTC ATT CCC CCT CAT ACT GTT TTT GGA TTC TCT CCA TCT GCT GCA TTT
ATC ACA ACG CTA AAG TAA GGG GGA GTA TGA CAA AAA CCT AAG AGA GGT AGA CGA CGT AAA

13681
GCA CCA TTA AAT GTA ATA TTA TTT GTT CCT GTA CCA CTG GTT CCA TTA GTA TTT CCT GTT
CGT GGT AAT TTA CAT TAT AAT AAA CAA GGA CAT GGT GAC CAA GGT AAT CAT AAA GGA CAA
13741
ACG ATG TTT CCA GTC ATG GTA GCA CCA CCA CTA AAA GTA GCA TTG AAT GTT GCG CCA GTA
TGC TAC AAA GGT CAG TAC CAT CGT GGT GGT GAT TTT CAT CGT AAC TTA CAA CGC GGT CAT

13801      ⇨ ntf034 ⇨
TTG ACA TTG ATT ACA GGC CCT CCG CTT GCG ATG TTT GTA TTT AGA GTT GCA TTT CCA GCG
AAC TGT AAC TAA TGT CCG GGA GGC GAA CGC TAC AAA CAT AAA TCT CAA CGT AAA GGT CGC

13861
CCA AGG ATG TTA AGA TTC CCA ATC AAA TTT TTT GCA TTA GCT TGA TTT GCA CCA CCA AGA
GGT TCC TAC AAT TCT AAG GGT TAG TTT AAA AAA CGT AAT CGA ACT AAA CGT GGT GGT TCT

13921
⇨ ✂ (Bgl II) ntf049 ⇨
TTC AGG GTG AAG GTT TTA GCA GGG CCA CTA CTA GCA ATC CCA CCA AAA TCA AAA TTC AAG
AAG TCC CAC TTC CAA AAT CGT CCC GGT GAT GAT CGT TAG GGT GGT TTT AGT TTT AAG TTC
                                ⇦ ntf053 ⇦
13981
TTT ACT GGC TTT CCT GCA GCG CCT AGA TTT TGA TTT GCT GGA TCT GCT GGG TCT GCA CCT
AAA TGA CCG AAA GGA CGT CGC GGA TCT AAA ACT AAA CGA CCT AGA CGA CCC AGA CGT GGA

14041
TGT GCA TTT CCA CCT GCA CCT ACT AGC TTG GCA TCC CCG TTG CAT TTG CGC TAT CGT TAT
ACA CGT AAA GGT GGA CGT GGA TGA TCG AAC CGT AGG GGC AAC GTA AAC GCG ATA GCA ATA
                                                  ⇦ ntf023 ⇦
14101
TAA ATT TTT GCA AAT AAT TTT GGA TCT GAT CAA ATC TTG CAA TAG CAT TAG CAT TAG CAT
ATT TAA AAA CGT TTA TTA AAA CCT AGA CTA GTT TAG AAC GTT ATC GTA ATC GTA ATC GTA

14161
TAG CAT TAT TGC CAC TAC CAA GCC CGC CTG CTT GTT CCT TCA AAT CTT TTA GAT TTG AAG
ATC GTA ATA ACG GTG ATG GTT CGG GCG GAC GAA CAA GGA AGT TTA GAA AAT CTA AAC TTC

14221
GAG GGT TGG AAA ACA GTG GCC TGA TTT GCT TGA CCT GCA CCA CCT GCT GCT GGT TTG ATT
CTC CCA ACC TTT TGT CAC CGG ACT AAA CGA ACT GGA CGT GGT GGA CGA CGA CCA AAC TAA

14281
GTC GCA TGA GTA AGG GTT GGA GAA TTA TTC ACT CCA TTG GTG TTA TTG AAT GTG ATT CCA
CAG CGT ACT CAT TCC CAA CCT CTT AAT AAG TGA GGT AAC CAC AAT AAC TTA CAC TAA GGT

14341
EcoRI
GAA TTC TGG ATG ATT AGA GTG TTT CCT CCA TTG GTG GTA ATC CAC AGT CAC ATT CCC AAA
CTT AAG ACC TAC TAA TCT CAC AAA GGA GGT AAC CAC CAT TAG GTG TCA GTG TAA GGG TTT


14401                                                  ⇨ ✂ (Bgl II) ntf049 ⇨
AAT CCC TAC CAG GCT TTT GCA TTA GCA TTA CCT GCA CCA CCA AGA TTC AGG GTG AAG GTT
TTA GGG ATG GTC CGA AAA CGT AAT CGT AAT GGA CGT GGT GGT TCT AAG TCC CAC TTC CAA

14461
TTA GCA GTG CCA CTA CTT GCA ATC CCA CCA AAA TCA AAA TTC AAG TTT ACT GGC TTT CCT
AAT CGT CAC GGT GAT GAA CGT TAG GGT GGT TTT AGT TTT AAG TTC AAA TGA CCG AAA GGA

14521
TCA GCG CCT AGA TTT TGA TTT GCT TGA GCT GCT GGT TTT CCA TCC ACC GTA AAA TTA TTC
AGT CGC GGA TCT AAA ACT AAA CGA ACT CGA CGA CCA AAA GGT AGG TGG CAT TTT AAT AAG

14581
ACA GTA TTT GCG ATG TTT ACT TCA ATG TTT TGT GCT GTT ACT GGT GCT GCG GGA TTT GTC
TGT CAT AAA CGC TAC AAA TGA AGT TAC AAA ACA CGA CAA TGA CCA CGA CGC CCT AAA CAG
```

⇦ ntf064 ⇦

```
14641
ACT GCT GTT CCA TTT ACA GTG GTG TTA GGC ATC ATA GGT AGC ATT GCC ACC AGT TCC ATC
TGA CGA CAA GGT AAA TGT CAC CAC AAT CCG TAG TAT CCA TCG TAA CGG TGG TCA AGG TAG

14701
AGG CGC GAC GCC GCT GTT AGT GCT AGA AGT TCC ATT ATT TTG GAA CAA AAG ATT ATT GGA
TCC GCG CTG CGG CGA CAA TCA CGA TCT TCA AGG TAA TAA AAC CTT GTT TTC TAA TAA CCT

14761                                          ⇨ ntf050 ⇨
TCC AGC CCA CCC TGG TGT GTT TGT TAC TAC GTT TCC AGT AAG CAA AAT GGG GTT AGT TAC
AGG TCG GGT GGG ACC ACA CAA ACA ATG ATG CAA AGG TCA TTC GTT TTA CCC CAA TCA ATG

14821
GAT GTT GTT TGT ATT CAT CAC AAG TGT GTT GAT TCC TTG GTC CAG TGT AGA CGT GAG ATT
CTA CAA CAA ACA TAA GTA GTG TTC ACA CAA CTA AGG AAC CAG GTC ACA TCT GCA CTC TAA

14881
GGC ATT CAC CAA ACC TCC ATC TGT GAA TAC GAT GTT ATT
CCG TAA GTG GTT TGG AGG TAG ACA CTT ATG CTA CAA TAA
```

APPENDIX 3

DNA sequences repeated in the *hsr* locus of *H. mustelae* strain 4298. The data was generated by DNA Strider (Dr Christian Marck, France) repeat search program for the entire 14919 bp DNA sequence of the HSRL. The criteria for the search were as follows: at least 12 nucleotides in length with 100% identity and no overlapping sequences were allowed. Repeats were grouped and numbered according to sequence identities, and the corresponding HSRL start and end coordinates are shown. The orientation of each repeat, with respect to the *hsr* gene, is represented by a + when the sequence is in the same orientation or – when in the opposite orientation to the gene. Descriptions of subtypes are given in sections 3.1 and 3.4.2, which are denoted with a single letter.

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
61	1	30	30	-	HU	4
44	1	36	36	-	U	3
68	9	34	26	-	U	4
84	15	37	23	-	UD	3
77	15	38	24	-	HUD	4
265	45	56	12	-	HU	2
224	70	82	13	-	U	2
266	72	83	12	-	UD	4
267	74	85	12	-	HU	2
164	231	245	15	-	HU	4
268	267	278	12	-	HU	2
225	267	279	13	-	U	2
269	268	279	12	-	HU	3
226	646	658	13	-	HU	2
92	722	743	22	-	HU	2
68	722	747	26	-	U	4
100	728	747	20	-	HUD	6
112	749	766	18	-	UD	2
141	770	785	16	-	UD	2
39	770	808	39	-	UD	2
113	791	808	18	-	HU	2
270	793	804	12	-	HUD	5
227	796	808	13	-	UD	3
188	796	809	14	-	U	2
271	797	808	12	-	U	6
142	802	817	16	-	U	2
106	802	820	19	-	U	2
189	805	818	14	-	HU	3
124	807	823	17	-	HUD	4
165	812	826	15	-	UD	5
114	812	829	18	-	UD	2
107	812	830	19	-	U	2
125	813	829	17	-	U	3
272	819	830	12	-	HU	4
228	820	832	13	-	U	3
190	820	833	14	-	HU	2
9	823	946	124	-	UD	2
53	842	875	34	-	HU	2
229	877	889	13	-	UD	3

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
21	879	947	69	-	HU	2
230	925	937	13	-	HUD	8
18	951	1024	74	-	UD	2
47	952	986	35	-	HU	3
115	973	990	18	-	UD	5
22	985	1052	68	-	HU	3
231	1040	1052	13	-	UD	2
116	1040	1057	18	-	U	3
273	1041	1052	12	-	HUD	11
126	1041	1057	17	-	UD	3
232	1045	1057	13	-	UD	2
127	1053	1069	17	-	U	2
66	1060	1087	28	-	HUD	4
17	1060	1135	76	-	U	2
85	1065	1087	23	-	U	3
128	1092	1108	17	-	U	3
274	1093	1104	12	-	U	5
32	1100	1143	44	-	UD	2
48	1101	1135	35	-	UD	3
35	1101	1142	42	-	HUD	4
108	1103	1121	19	-	HUD	7
93	1145	1166	22	-	HUD	4
10	1145	1268	124	-	UD	2
275	1154	1165	12	-	UD	5
328	1156	1167	12	-	UD	3
331	1157	1168	12	-	HUD	4
117	1157	1174	18	-	UD	3
166	1158	1172	15	-	U	2
277	1159	1170	12	-	HUD	8
227	1162	1174	13	-	UD	3
191	1162	1175	14	-	HUD	4
233	1163	1175	13	-	HUD	8
167	1163	1177	15	-	UD	4
278	1166	1177	12	-	UD	4
235	1166	1178	13	-	HUD	5
234	1171	1183	13	-	HUD	6
279	1190	1201	12	-	UD	3
78	1204	1227	24	-	UD	4
24	1213	1279	67	-	U	3
236	1229	1241	13	-	HUD	6
101	1246	1265	20	-	UD	5
168	1284	1298	15	-	HU	2
229	1293	1305	13	-	UD	3
237	1311	1323	13	-	U	3
280	1325	1336	12	-	HUD	3
3	1337	1638	302	-	U	2
274	1350	1361	12	-	U	5
143	1372	1387	16	-	U	4
64	1380	1408	29	-	U	4
58	1397	1429	33	-	HU	3
164	1404	1418	15	-	HU	4
192	1410	1423	14	-	HU	5
144	1416	1431	16	-	U	3
238	1417	1429	13	-	U	4
102	1417	1436	20	-	U	3
239	1432	1444	13	-	U	5
193	1432	1445	14	-	U	5
103	1432	1451	20	-	U	3
282	1433	1444	12	-	U	6
194	1438	1451	14	-	U	2
330	1439	1450	12	-	U	6
240	1445	1457	13	-	UD	3
241	1451	1463	13	-	U	3
145	1451	1466	16	-	U	3
195	1459	1475	14	-	HU	5

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
334	1505	1518	14	+	UD	3
30	1533	1580	48	-	HU	3
129	1542	1558	17	-	HU	4
242	1560	1572	13	-	HU	5
284	1561	1572	12	-	HU	5
235	1561	1573	13	-	HUD	5
130	1564	1580	17	-	U	4
170	1567	1581	15	-	U	3
131	1577	1593	17	-	UD	3
146	1583	1598	16	-	HU	3
285	1591	1602	12	-	U	5
286	1593	1604	12	-	HUD	7
165	1593	1607	15	-	UD	5
196	1594	1607	14	-	U	3
243	1598	1610	13	-	U	4
197	1612	1625	14	-	HUD	7
171	1612	1626	15	-	U	3
287	1630	1641	12	-	UD	3
132	1638	1654	17	-	HU	2
198	1641	1654	14	-	U	2
147	1643	1658	16	-	U	3
2	1656	2023	368	-	U	2
143	1663	1678	16	-	U	4
118	1688	1705	18	-	U	3
59	1688	1719	32	-	HU	3
109	1689	1707	19	-	HU	4
119	1690	1707	18	-	HU	5
244	1693	1705	13	-	HU	7
11	1710	1831	122	-	UD	3
6	1710	1846	137	-	HU	3
288	1741	1752	12	-	HUD	5
339*	1789	1802	14	+	HUD	4
29	1832	1882	51	-	UD	3
266	1865	1876	12	-	UD	4
72	1870	1864	25	-	UD	3
133	1874	1890	17	-	HUD	4
245	1878	1890	13	-	UD	5
16	1879	1962	84	-	HU	3
54	1904	1937	34	-	HU	5
230	1921	1933	13	-	HUD	8
329	1925	1936	12	-	HU	6
148	1930	1945	16	-	HU	4
325	1934	1945	12	-	HU	5
199	1955	1968	14	-	U	3
149	1955	1970	16	-	U	4
242	1966	1978	13	-	HU	5
278	1967	1978	12	-	UD	4
200	1967	1980	14	-	U	3
150	1985	2000	16	-	U	3
119	1985	2002	18	-	HU	5
73	1985	2009	25	-	U	3
244	1988	2000	13	-	HU	7
246	2011	2023	13	-	HU	2
172	2011	2025	15	-	U	2
102	2011	2030	20	-	U	3
291	2025	2036	12	-	U	3
173	2025	2039	15	-	U	2
282	2036	2047	12	-	U	6
281	2041	2052	12	-	U	3
239	2044	2056	13	-	U	5
281	2050	2061	12	-	U	3
193	2053	2066	14	-	U	5
282	2054	2065	12	-	U	6
142	2059	2074	16	-	U	2
247	2062	2074	13	-	HU	2

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
292	2064	2075	12	-	HU	2
151	2069	2084	16	-	U	2
239	2071	2083	13	-	U	5
193	2071	2084	14	-	U	5
282	2072	2083	12	-	U	6
106	2077	2095	19	-	U	2
189	2080	2093	14	-	HU	3
201	2082	2095	14	-	HUD	4
293	2087	2098	12	-	U	3
248	2088	2100	13	-	U	2
294	2089	2100	12	-	HU	5
152	2089	2104	16	-	HU	3
1	2089	2829	741	-	U	2
295	2090	2101	12	-	HU	4
296	2091	2102	12	-	HU	3
228	2092	2104	13	-	U	3
174	2105	2119	15	-	U	4
28	2105	2160	56	-	UD	3
24	2105	2171	67	-	U	3
236	2121	2133	13	-	HUD	6
101	2138	2157	20	-	UD	5
130	2163	2179	17	-	U	4
65	2163	2191	29	-	HU	3
202	2166	2179	14	-	U	3
168	2176	2190	15	-	HU	2
243	2197	2209	13	-	U	4
94	2203	2224	22	-	HU	3
197	2211	2224	14	-	HUD	7
249	2225	2237	13	-	U	3
297	2226	2237	12	-	U	3
298	2239	2250	12	-	HU	4
147	2239	2254	16	-	U	3
64	2260	2288	29	-	U	4
299	2277	2288	12	-	HU	3
192	2290	2303	14	-	HU	5
300	2315	2326	12	-	HU	4
14	2335	2420	86	-	HU	3
195	2336	2349	14	-	HU	5
203	2407	2420	14	-	UD	3
301	2409	2420	12	-	HUD	5
204	2426	2439	14	-	UD	4
79	2450	2473	24	-	U	3
234	2453	2465	13	-	HUD	6
42	2476	2512	37	-	U	3
250	2480	2492	13	-	U	4
205	2492	2505	14	-	HUD	5
115	2492	2509	18	-	UD	5
302	2501	2512	12	-	HU	4
175	2514	2528	15	-	U	3
303	2517	2528	12	-	HU	3
134	2531	2547	17	-	U	2
277	2534	2545	12	-	HUD	8
135	2540	2556	17	-	U	2
277	2543	2554	12	-	HUD	8
305	2549	2560	12	-	UD	7
120	2591	2608	18	-	UD	3
251	2593	2605	13	-	HUD	4
305	2617	2628	12	-	UD	7
61	2620	2649	30	-	HU	4
44	2620	2655	36	-	U	3
68	2628	2653	26	-	U	4
95	2634	2655	22	-	HUD	7
153	2659	2674	16	-	HUD	3
86	2675	2697	23	-	HU	3
305	2699	2710	12	-	UD	7

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
330	2709	2720	12	-	U	6
252	2711	2723	13	-	U	5
293	2715	2726	12	-	U	3
252	2717	2729	13	-	U	5
55	2717	2750	34	-	U	3
285	2719	2730	12	-	U	5
176	2727	2741	15	-	HU	4
154	2741	2756	16	-	UD	3
80	2747	2770	24	-	U	3
149	2757	2772	16	-	U	4
326	2767	2778	12	-	HU	3
270	2774	2785	12	-	HUD	5
136	2787	2803	17	-	UD	3
253	2792	2804	13	-	U	3
13	2792	2882	91	-	U	2
155	2821	2836	16	-	U	3
271	2824	2835	12	-	U	6
233	2824	2836	13	-	HUD	8
167	2824	2838	15	-	UD	4
284	2827	2838	12	-	HU	5
200	2827	2840	14	-	U	3
206	2841	2854	14	-	U	3
78	2847	2870	24	-	UD	4
174	2856	2870	15	-	U	4
254	2861	2853	13	-	U	6
87	2864	2886	23	-	HU	2
309	2865	2876	12	-	HU	4
236	2872	2884	13	-	HUD	6
101	2889	2908	20	-	UD	5
202	2917	2930	14	-	U	3
170	2917	2931	15	-	U	3
307	2925	2936	12	-	UD	3
207	2925	2938	14	-	UD	3
156	2925	2940	16	-	UD	2
96	2927	2948	22	-	UD	3
255	2928	2940	13	-	HU	4
116	2928	2945	18	-	U	3
273	2929	2940	12	-	HUD	11
121	2929	2946	18	-	UD	3
104	2929	2948	20	-	UD	2
208	2933	2946	14	-	UD	3
273	2935	2946	12	-	HUD	11
209	2937	2950	14	-	HU	2
177	2946	2960	15	-	UD	2
306	2949	2960	12	-	HUD	4
237	2951	2963	13	-	U	3
137	2951	2967	17	-	U	2
122	2955	2972	18	-	UD	2
197	2959	2972	14	-	HUD	7
171	2959	2973	15	-	U	3
3	2977	3278	302	-	U	2
274	2990	3001	12	-	U	5
143	3012	3027	16	-	U	4
64	3020	3048	29	-	U	4
58	3037	3069	33	-	HU	3
164	3044	3058	15	-	HU	4
192	3050	3063	14	-	HU	5
144	3056	3071	16	-	U	3
238	3057	3069	13	-	U	4
102	3057	3076	20	-	U	3
239	3072	3084	13	-	U	5
193	3072	3085	14	-	U	5
103	3072	3091	20	-	U	3
282	3073	3084	12	-	U	6
330	3079	3090	12	-	U	6

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
240	3085	3097	13	-	UD	3
241	3091	3103	13	-	U	3
145	3091	3106	16	-	U	3
195	3099	3112	14	-	HU	5
334	3145	3158	14	+	UD	3
30	3173	3220	48	-	HU	3
129	3182	3198	17	-	HU	4
242	3200	3212	13	-	HU	5
284	3201	3212	12	-	HU	5
235	3201	3213	13	-	HUD	5
130	3204	3220	17	-	U	4
170	3207	3221	15	-	U	3
131	3217	3233	17	-	UD	3
146	3223	3238	16	-	HU	3
285	3231	3242	12	-	U	5
286	3233	3244	12	-	HUD	7
165	3233	3247	15	-	UD	5
196	3234	3247	14	-	U	3
243	3238	3250	13	-	U	4
197	3252	3265	14	-	HUD	7
171	3252	3266	15	-	U	3
249	3269	3281	13	-	U	3
91	3270	3291	22	-	U	2
198	3281	3294	14	-	U	2
98	3281	3301	21	-	HU	2
298	3283	3294	12	-	HU	4
2	3296	3663	368	-	U	2
143	3303	3318	16	-	U	4
118	3328	3345	18	-	U	3
59	3328	3359	32	-	HU	3
109	3329	3347	19	-	HU	4
119	3330	3347	18	-	HU	5
244	3333	3345	13	-	HU	7
11	3350	3471	122	-	UD	3
6	3350	3486	137	-	HU	3
288	3381	3392	12	-	HUD	5
339*	3429	3442	14	+	HUD	4
29	3472	3522	51	-	UD	3
266	3505	3516	12	-	UD	4
72	3510	3534	25	-	UD	3
133	3514	3530	17	-	HUD	4
245	3518	3530	13	-	UD	5
16	3519	3602	84	-	HU	3
54	3544	3577	34	-	HU	5
230	3561	3573	13	-	HUD	8
329	3565	3576	12	-	HU	6
148	3570	3585	16	-	HU	4
325	3574	3585	12	-	HU	5
199	3595	3608	14	-	U	3
149	3595	3610	16	-	U	4
242	3606	3618	13	-	HU	5
278	3607	3618	12	-	UD	4
150	3625	3640	16	-	U	3
119	3625	3642	18	-	HU	5
73	3625	3649	25	-	U	3
244	3628	3640	13	-	HU	7
238	3651	3663	13	-	U	4
210	3654	3664	14	-	HU	2
178	3659	3673	15	-	U	2
179	3663	3677	15	-	U	2
180	3667	3681	15	-	U	2
316	3672	3683	12	-	UD	2
291	3674	3685	12	-	U	3
173	3674	3688	15	-	U	2
155	3682	3697	16	-	U	3

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
271	3685	3696	12	-	U	6
233	3685	3697	13	-	HUD	8
158	3685	3700	16	-	HUD	3
281	3690	3701	12	-	U	3
224	3692	3704	13	-	U	2
181	3694	3708	15	-	HU	2
331	3697	3708	12	-	HUD	4
117	3697	3714	18	-	UD	3
166	3698	3712	15	-	U	2
277	3699	3710	12	-	HUD	8
256	3702	3714	13	-	HUD	3
188	3702	3715	14	-	U	2
271	3703	3714	12	-	U	6
194	3708	3721	14	-	U	2
182	3708	3722	15	-	U	2
330	3709	3720	12	-	U	6
294	3711	3722	12	-	HU	5
152	3711	3726	16	-	HU	3
1	3711	4451	741	-	U	2
295	3712	3723	12	-	HU	4
296	3713	3724	12	-	HU	3
228	3714	3726	13	-	U	3
174	3727	3741	15	-	U	4
28	3727	3782	56	-	UD	3
24	3727	3793	67	-	U	3
236	3743	3755	13	-	HUD	6
101	3760	3779	20	-	UD	5
130	3785	3801	17	-	U	4
65	3785	3813	29	-	HU	3
202	3788	3801	14	-	U	3
186	3798	3812	15	-	HU	2
243	3819	3831	13	-	U	4
94	3825	3846	22	-	HU	3
197	3833	3846	14	-	HUD	7
249	3847	3859	13	-	U	3
297	3848	3859	12	-	U	3
298	3861	3872	12	-	HU	4
147	3861	3876	16	-	U	3
64	3882	3910	29	-	U	4
299	3899	3910	12	-	HU	3
192	3912	3925	14	-	HU	5
300	3937	3948	12	-	HU	4
14	3957	4042	86	-	HU	3
195	3958	3971	14	-	HU	5
203	4029	1042	14	-	UD	3
301	4031	4042	12	-	HUD	5
204	4048	4061	14	-	UD	4
79	4072	4095	24	-	U	3
234	4075	4087	13	-	HUD	6
42	4098	4134	37	-	U	3
250	4102	4114	13	-	U	4
205	4114	4127	14	-	HUD	5
115	4114	4131	18	-	UD	5
302	4123	4134	12	-	HU	4
175	4136	4150	15	-	U	3
303	4139	4150	12	-	HU	3
135	4153	4169	17	-	U	2
277	4156	4167	12	-	HUD	8
290	4161	4172	12	-	HU	2
134	4162	4178	17	-	U	2
277	4165	4176	12	-	HUD	8
305	4171	4182	12	-	UD	7
120	4213	4230	18	-	UD	3
251	4215	4227	13	-	HUD	4
305	4239	4250	12	-	UD	7

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
61	4242	4271	30	-	HU	4
44	4242	4277	36	-	U	3
68	4250	4275	26	-	U	4
95	4256	4277	22	-	HUD	7
153	4281	4296	16	-	HUD	3
86	4297	4319	23	-	HU	3
305	4321	4332	12	-	UD	7
330	4331	4342	12	-	U	6
252	4333	4345	13	-	U	5
293	4337	4348	12	-	U	3
252	4339	4351	13	-	U	5
55	4339	4372	34	-	U	3
285	4341	4352	12	-	U	5
176	4349	4363	15	-	HU	4
154	4363	4378	16	-	UD	3
80	4369	4392	24	-	U	3
149	4379	4394	16	-	U	4
326	4389	4400	12	-	HU	3
270	4396	4407	12	-	HUD	5
136	4409	4425	17	-	UD	3
253	4414	4426	13	-	U	3
41	4414	4451	38	-	U	2
20	4455	4526	72	-	U	2
245	4456	4468	13	-	UD	5
160	4456	4471	16	-	HU	3
318	4457	4468	12	-	HU	4
264	4474	4485	12	-	HUD	5
54	4485	4518	34	-	HU	5
230	4502	4514	13	-	HUD	8
329	4506	4517	12	-	HU	6
257	4508	4520	13	-	HU	3
324	4512	4523	12	-	HU	3
323	4551	4562	12	-	UD	2
179	4553	4567	15	-	U	2
244	4572	4584	13	-	HU	7
184	4572	4586	15	-	U	2
253	4574	4585	13	-	U	3
322	4574	4585	12	-	UD	3
180	4584	4598	15	-	U	2
178	4585	4599	15	-	U	2
271	4593	4604	12	-	U	6
233	4593	4605	13	-	HUD	8
12	4601	4702	102	-	UD	2
34	4612	4654	43	-	HUD	3
264	4624	4635	12	-	HUD	5
15	4644	4728	85	-	UD	2
258	4651	4663	13	-	HUD	4
40	4698	4736	39	-	HU	2
69	4703	4728	26	-	HUD	5
26	4703	4761	59	-	UD	2
161	4706	4721	16	-	HUD	6
100	4706	4725	20	-	HUD	6
95	4706	4727	22	-	HUD	7
77	4706	4729	24	-	HUD	4
81	4738	4761	24	-	HU	2
111	4793	4810	18	-	HU	2
272	4799	4810	12	-	HU	4
162	4800	4815	16	-	HU	2
300	4803	4814	12	-	HU	4
225	4804	4816	13	-	U	2
269	4805	4816	12	-	HU	3
252	4806	4818	13	-	U	5
55	4806	4839	34	-	U	3
285	4808	4819	12	-	U	5
176	4816	4830	15	-	HU	4

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
80	4847	4870	24	-	U	3
199	4857	4870	14	-	U	3
328	4862	4873	12	-	UD	3
74	4878	4902	25	-	HU	2
118	4885	4902	18	-	U	3
138	4886	4902	17	-	U	2
150	4887	4902	16	-	U	3
184	4890	4904	15	-	U	2
322	4892	4903	12	-	UD	3
41	4892	4929	38	-	U	2
13	4892	4982	91	-	U	2
155	4921	4936	16	-	U	3
271	4924	4935	12	-	U	6
233	4924	4936	13	-	HUD	8
167	4924	4938	15	-	UD	4
284	4927	4938	12	-	HU	5
200	4927	4940	14	-	U	3
254	4941	4953	13	-	U	6
206	4941	4954	14	-	U	3
78	4947	4970	24	-	UD	4
174	4956	4970	15	-	U	4
110	4964	4982	19	-	HU	2
309	4965	4976	12	-	HU	4
129	4998	5014	17	-	HU	4
97	5017	5037	21	-	UD	2
321	5033	5044	12	-	U	2
211	5033	5046	14	-	UD	2
96	5033	5054	22	-	UD	3
255	5034	5046	13	-	HU	4
116	5034	5051	18	-	U	3
273	5035	5046	12	-	HUD	11
121	5035	5052	18	-	UD	3
75	5035	5059	25	-	UD	2
208	5039	5052	14	-	UD	3
273	5041	5052	12	-	HUD	11
320	5043	5054	12	-	HUD	3
237	5057	5069	13	-	U	3
137	5057	5073	17	-	U	2
260	5061	5073	13	-	UD	2
297	5080	5091	12	-	U	3
91	5080	5101	22	-	U	2
301	5128	5139	12	-	HUD	5
261	5137	5149	13	-	HU	2
89	5138	5160	23	-	UD	2
286	5143	5154	12	-	HUD	7
165	5143	5157	15	-	UD	5
107	5143	5161	19	-	U	2
125	5144	5160	17	-	U	3
272	5150	5161	12	-	HU	4
33	5150	5193	44	-	HU	2
76	5169	5193	25	-	UD	2
212	5199	5212	14	-	HU	2
279	5238	5249	12	-	UD	3
127	5239	5255	17	-	U	2
37	5251	5291	41	-	UD	2
287	5272	5283	12	-	UD	3
207	5280	5293	14	-	UD	3
321	5282	5293	12	-	U	2
319	5295	5306	12	-	U	2
85	5296	5318	23	-	U	3
67	5296	5323	28	-	UD	2
19	5296	5369	74	-	HUD	3
48	5335	5369	35	-	UD	3
36	5335	5376	42	-	U	2
108	5337	5355	19	-	HUD	7

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
241	5364	5376	13	-	U	3
62	5371	5400	30	-	UD	2
327	5378	5389	12	-	HUD	5
70	5378	5403	26	-	HUD	3
93	5379	5400	22	-	HUD	4
275	5388	5399	12	-	UD	5
139	5392	5408	17	-	UD	2
204	5393	5406	14	-	UD	4
262	5397	5409	13	-	UD	2
151	5403	5418	16	-	U	2
239	5405	5417	13	-	U	5
193	5405	5418	14	-	U	5
103	5405	5424	20	-	U	3
282	5406	5417	12	-	U	6
182	5411	5425	15	-	U	2
330	5412	5423	12	-	U	6
294	5414	5425	12	-	HU	5
263	5414	5426	13	-	U	2
79	5417	5440	24	-	U	3
234	5420	5432	13	-	HUD	6
185	5437	5451	15	-	HUD	3
42	5443	5479	37	-	U	3
250	5447	5459	13	-	U	4
205	5459	5472	14	-	HUD	5
115	5459	5476	18	-	UD	5
302	5468	5479	12	-	HU	4
175	5481	5495	15	-	U	3
183	5484	5498	15	-	HU	2
196	5503	5516	14	-	U	3
125	5503	5519	17	-	U	3
218	5510	5522	13	-	UD	3
291	5515	5526	12	-	U	3
20	5526	5597	72	-	U	2
245	5527	5539	13	-	UD	5
160	5527	5542	16	-	HU	3
318	5528	5539	12	-	HU	4
264	5545	5556	12	-	HUD	5
54	5556	5589	34	-	HU	5
230	5573	5585	13	-	HUD	8
329	5577	5588	12	-	HU	6
257	5579	5591	13	-	HU	3
324	5583	5594	12	-	HU	3
317	5598	5609	12	-	HU	2
275	5610	5621	12	-	UD	5
315	5613	5624	12	-	UD	2
295	5619	5630	12	-	HU	4
218	5621	5633	13	-	UD	3
138	5636	5652	17	-	U	2
109	5636	5654	19	-	HU	4
73	5637	5661	25	-	U	3
244	5640	5652	13	-	HU	7
213	5662	5675	14	-	HU	2
144	5662	5677	16	-	U	3
238	5663	5675	13	-	U	4
172	5663	5677	15	-	U	2
214	5691	5704	14	-	U	2
90	5691	5713	23	-	U	2
254	5692	5704	13	-	U	6
254	5701	5713	13	-	U	6
254	5710	5722	13	-	U	6
90	5718	5740	23	-	U	2
254	5719	5731	13	-	U	6
214	5727	5740	14	-	U	2
206	5728	5741	14	-	U	3
314	6317	6328	12	-	UD	2

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
223	6472	6484	13	-	U	2
5	6517	6670	154	-	HU	2
47	6526	6560	35	-	HU	3
49	6526	6560	35	-	UD	2
205	6547	6560	14	-	HUD	5
313	6563	6574	12	-	HUD	4
38	6571	6610	40	-	HUD	3
22	6571	6638	68	-	HU	3
255	6626	6638	13	-	HU	4
31	6626	6670	45	-	UD	2
273	6627	6638	12	-	HUD	11
312	6633	6644	12	-	HUD	4
201	6633	6646	14	-	HUD	4
124	6633	6649	17	-	HUD	4
286	6638	6649	12	-	HUD	7
306	6641	6652	12	-	HUD	4
63	6641	6670	30	-	UD	2
17	6643	6718	76	-	U	2
85	6648	6670	23	-	U	3
128	6675	6691	17	-	U	3
274	6676	6687	12	-	U	5
45	6683	6718	36	-	UD	2
50	6684	6718	35	-	HUD	3
36	6684	6725	42	-	U	2
108	6686	6704	19	-	HUD	7
145	6713	6728	16	-	U	3
159	6720	6735	16	-	HUD	3
248	6738	6750	13	-	U	2
294	6739	6750	12	-	HU	5
263	6739	6751	13	-	U	2
99	6745	6764	20	-	UD	2
105	6746	6764	19	-	HU	2
250	6766	6778	13	-	U	4
140	6772	6787	16	-	HU	2
319	6791	6802	12	-	U	2
123	7038	7054	17	-	U	2
123	7125	7141	17	-	U	2
338	7263	7278	16	+	H	2
128	7272	7288	17	-	U	3
274	7273	7284	12	-	U	5
335	7370	7382	13	+	UD	2
223	7381	7393	13	-	U	2
311	7393	7404	12	-	U	2
311	7432	7443	12	-	U	2
169	7486	7500	15	+	HD	2
6	7487	7623	137	+	HU	3
8	7502	7628	127	+	HD	2
339*	7531	7544	14	+	HUD	4
288	7581	7592	12	+	HUD	5
310	7615	7626	12	+	HD	3
187	7615	7628	14	+	HD	2
189	7627	7640	14	+	HU	3
312	7628	7639	12	+	HUD	4
247	7628	7640	13	+	HU	2
289	7636	7647	12	+	HD	2
261	7636	7648	13	+	HU	2
59	7653	7684	32	+	HU	3
119	7665	7682	18	+	HU	5
109	7665	7683	19	+	HU	4
244	7666	7678	13	+	HU	7
74	7667	7691	25	+	HU	2
221	7682	7694	13	+	H	3
16	7696	7779	84	+	HU	3
325	7713	7724	12	+	HU	5
148	7713	7728	16	+	HU	4

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
54	7721	7754	34	+	HU	5
329	7722	7733	12	+	HU	6
230	7725	7737	13	+	HUD	8
157	7764	7779	16	+	HD	2
318	7768	7779	12	+	HU	4
234	7775	7787	13	+	HUD	6
326	7789	7800	12	+	HU	3
25	7799	7859	61	+	HD	2
140	7818	7833	16	+	HU	2
333*	7818	7843	26	+	HD	2
185	7834	7848	15	+	HUD	3
105	7841	7859	19	+	HU	2
181	7853	7867	15	+	HU	2
331	7854	7865	12	+	HUD	4
221	7859	7871	13	+	H	3
304	7861	7872	12	+	HD	3
70	7862	7887	26	+	HUD	3
7	7862	7991	130	+	HD	2
93	7865	7886	22	+	HUD	4
88	7865	7887	23	+	HD	3
327	7876	7887	12	+	HUD	5
159	7879	7894	16	+	HUD	3
35	7889	7930	42	+	HUD	4
50	7896	7930	35	+	HUD	3
19	7896	7968	74	+	HUD	3
108	7910	7928	19	+	HUD	7
338	7915	7930	16	+	H	2
52	7942	7976	35	+	HD	3
66	7947	7974	28	+	HUD	4
5	7947	8100	154	+	HU	2
306	7964	7975	12	+	HUD	4
286	7967	7978	12	+	HUD	7
124	7968	7984	17	+	HUD	4
201	7970	7983	14	+	HUD	4
312	7973	7984	12	+	HUD	4
255	7978	7990	13	+	HU	4
273	7979	7990	12	+	HUD	11
22	7979	8046	68	+	HU	3
38	8007	8046	40	+	HUD	3
313	8043	8054	12	+	HUD	4
205	8056	8069	14	+	HUD	5
51	8057	8091	35	+	HD	2
47	8057	8091	35	+	HU	3
23	8079	8164	68	+	HD	2
21	8096	8164	69	+	HU	2
230	8106	8118	13	+	HUD	8
57	8168	8201	34	+	HD	2
53	8168	8201	34	+	HU	2
176	8192	8206	15	+	HU	4
34	8209	8251	43	+	HUD	3
264	8227	8238	12	+	HUD	5
318	8245	8256	12	+	HU	4
133	8245	8261	17	+	HUD	4
160	8252	8267	16	+	HU	3
71	8260	8285	26	+	HD	2
158	8264	8279	16	+	HUD	3
233	8266	8278	13	+	HUD	8
256	8267	8279	13	+	HUD	3
191	8267	8280	14	+	HUD	4
113	8268	8285	18	+	HU	2
270	8271	8282	12	+	HUD	5
86	8279	8301	23	+	HU	3
153	8291	8306	16	+	HUD	3
292	8362	8373	12	+	HU	2
309	8370	8381	12	+	HU	4

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
216	8422	8435	14	+	HD	2
308	8462	8473	12	+	HD	3
329	8487	8498	12	+	HU	6
340*	8533	8546	14	+	HD	2
313	8559	8570	12	+	HUD	4
308	8567	8578	12	+	HD	3
33	8583	8626	44	+	HU	2
272	8614	8625	12	+	HU	4
111	8615	8632	18	+	HU	2
163	8626	8640	15	+	HD	2
277	8629	8640	12	+	HUD	8
82	8633	8656	24	+	HD	2
301	8642	8653	12	+	HUD	5
14	8643	8728	86	+	HU	3
195	8713	8726	14	+	HU	5
212	8732	8745	14	+	HU	2
290	8740	8751	12	+	HU	2
210	8746	8759	14	+	HU	2
246	8747	8759	13	+	HU	2
213	8747	8760	14	+	HU	2
58	8748	8780	33	+	HU	3
192	8753	8766	14	+	HU	5
164	8758	8772	15	+	HU	4
299	8768	8779	12	+	HU	3
4	8790	8955	166	+	HD	2
81	8932	8955	24	+	HU	2
56	8957	8990	34	+	HD	2
40	8957	8995	39	+	HU	2
77	8964	8987	24	+	HUD	4
69	8965	8990	26	+	HUD	5
60	8965	8995	31	+	HD	2
95	8966	8987	22	+	HUD	7
100	8968	8987	20	+	HUD	6
161	8972	8987	16	+	HUD	6
257	8990	9002	13	+	HU	3
258	8996	9008	13	+	HUD	4
267	9012	9023	12	+	HU	2
217	9015	9028	14	+	HD	2
226	9029	9041	13	+	HU	2
98	9087	9107	21	+	HU	2
298	9093	9104	12	+	HU	4
132	9094	9110	17	+	HU	2
43	9100	9136	37	+	HD	2
341*	9110	9121	12	+	HD	2
280	9118	9129	12	+	HUD	3
197	9122	9135	14	+	HUD	7
94	9123	9144	22	+	HU	3
310	9136	9147	12	+	HD	3
190	9136	9149	14	+	HU	2
152	9138	9153	16	+	HU	3
295	9140	9151	12	+	HU	4
294	9141	9152	12	+	HU	5
268	9143	9154	12	+	HU	2
300	9144	9155	12	+	HU	4
162	9144	9159	16	+	HU	2
296	9148	9159	12	+	HU	3
146	9153	9168	16	+	HU	3
336*	9164	9176	13	+	HD	2
186	9171	9185	15	+	HU	2
65	9171	9199	29	+	HU	3
30	9183	9230	48	+	HU	3
235	9189	9201	13	+	HUD	5
284	9190	9201	12	+	HU	5
242	9190	9202	13	+	HU	5
332*	9194	9230	37	+	HD	2

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
129	9205	9221	17	+	HU	4
87	9227	9249	23	+	HU	2
236	9228	9240	13	+	HUD	6
110	9231	9249	19	+	HU	2
309	9237	9248	12	+	HU	4
288	9300	9311	12	+	HUD	5
265	9350	9361	12	+	HU	2
148	9480	9495	16	+	HU	4
327	9598	9609	12	+	HUD	5
161	9793	9808	16	+	HUD	6
92	9818	9839	22	+	HU	2
61	9818	9847	30	+	HU	4
324	9839	9850	12	+	HU	3
183	9920	9934	15	+	HU	2
303	9923	9934	12	+	HU	3
342*	10018	10029	12	+	HD	2
325	10131	10142	12	+	HU	5
325	10146	10157	12	+	HU	5
221	10199	10211	13	+	H	3
209	10770	10783	14	+	HU	2
320	10772	10783	12	+	HUD	3
222	10852	10864	13	+	H	2
251	10863	10875	13	+	HUD	4
222	10926	10938	13	+	H	2
302	11001	11012	12	+	HU	4
269	11209	11220	12	+	HU	3
337*	11864	11875	12	+	HD	2
335	12041	12053	13	+	UD	2
317	12081	12092	12	+	HU	2
215	12118	12131	14	+	HD	2
334	12119	12132	14	+	UD	3
219	12388	12400	13	-	D	2
337*	12422	12433	12	+	HD	2
283	12559	12570	12	-	D	2
343#	12571	12592	22	+	D	2
343#	12571	12592	22	-	D	2
336*	12686	12698	13	+	HD	2
342*	12751	12762	12	+	HD	2
108	12814	12832	19	-	HUD	7
219	12827	12839	13	-	D	2
216	12834	12847	14	-	HD	2
333*	12901	12926	26	+	HD	2
332*	12937	12973	37	+	HD	2
215	12988	13001	14	-	HD	2
154	13080	13095	16	-	UD	3
136	13105	13121	17	-	UD	3
322	13110	13121	12	-	UD	3
259	13123	13135	13	-	D	2
8	13123	13249	127	-	HD	2
310	13125	13136	12	-	HD	3
11	13128	13249	122	-	UD	3
288	13159	13170	12	-	HUD	5
339*	13207	13220	14	+	HUD	4
169	13249	13263	15	-	HD	2
29	13249	13299	51	-	UD	3
266	13282	13293	12	-	UD	4
220	13287	13299	13	-	D	2
217	13326	13339	14	-	HD	2
27	13334	13392	59	-	D	2
15	13334	13418	85	-	UD	2
258	13341	13353	13	-	HUD	4
60	13388	13418	31	-	HD	2
69	13393	13418	26	-	HUD	5
161	13396	13411	16	-	HUD	6
100	13396	13415	20	-	HUD	6

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
95	13396	13417	22	-	HUD	7
84	13396	13418	23	-	UD	3
112	13417	13434	18	-	UD	2
83	13427	13450	24	-	D	2
39	13435	13473	39	-	UD	2
71	13456	13481	26	-	HD	2
270	13458	13469	12	-	HUD	5
256	13461	13473	13	-	HUD	3
191	13461	13474	14	-	HUD	4
233	13462	13474	13	-	HUD	8
158	13462	13477	16	-	HUD	3
276	13471	13482	12	-	D	2
262	13471	13483	13	-	UD	2
240	13477	13489	13	-	UD	3
12	13479	13580	102	-	UD	2
34	13490	13532	43	-	HUD	3
264	13502	13513	12	-	HUD	5
27	13522	13580	59	-	D	2
258	13529	13529	13	-	HUD	4
305	13573	13584	12	-	UD	7
46	13577	13612	36	-	D	2
69	13587	13612	26	-	HUD	5
161	13590	13605	16	-	HUD	6
100	13590	13609	20	-	HUD	6
95	13590	13611	22	-	HUD	7
84	13590	13612	23	-	UD	3
83	13620	13643	24	-	D	2
141	13628	13643	16	-	UD	2
187	13667	13680	14	-	HD	2
259	13668	13680	13	-	D	2
9	13670	13793	124	-	UD	2
57	13689	13722	34	-	HD	2
229	13724	13736	13	-	UD	3
23	13726	13793	68	-	HD	2
230	13772	13784	13	-	HUD	8
18	13798	13871	74	-	UD	2
51	13799	13833	35	-	HD	2
49	13799	13833	35	-	UD	2
115	13820	13837	18	-	UD	5
38	13832	13871	40	-	HUD	3
37	13861	13901	41	-	UD	2
287	13882	13893	12	-	UD	3
307	13890	13901	12	-	UD	3
315	13902	13913	12	-	UD	2
177	13905	13919	15	-	UD	2
63	13908	13937	30	-	UD	2
52	13908	13942	35	-	HD	3
66	13910	13937	28	-	HUD	4
67	13915	13942	28	-	UD	2
45	13953	13988	36	-	UD	2
32	13953	13996	44	-	UD	2
48	13954	13988	35	-	UD	3
35	13954	13995	42	-	HUD	4
108	13956	13974	19	-	HUD	7
62	13990	14019	30	-	UD	2
327	13997	14008	12	-	HUD	5
88	13997	14019	23	-	HD	3
10	13998	14121	124	-	UD	2
275	14007	14018	12	-	UD	5
328	14009	14020	12	-	UD	3
331	14010	14021	12	-	HUD	4
117	14010	14027	18	-	UD	3
163	14011	14025	15	-	HD	2
277	14012	14023	12	-	HUD	8
227	14015	14027	13	-	UD	3

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
191	14015	14028	14	-	HUD	4
233	14016	14028	13	-	HUD	8
167	14016	14030	15	-	UD	4
278	14019	14030	12	-	UD	4
235	14019	14031	13	-	HUD	5
234	14024	14036	13	-	HUD	6
279	14043	14054	12	-	UD	3
78	14057	14080	24	-	UD	4
28	14066	14121	56	-	UD	3
236	14082	14094	13	-	HUD	6
101	14099	14118	20	-	UD	5
97	14121	14141	21	-	UD	2
283	14135	14146	12	-	D	2
131	14137	14153	17	-	UD	3
232	14143	14155	13	-	UD	2
208	14143	14156	14	-	UD	3
273	14145	14156	12	-	HUD	11
126	14145	14161	17	-	UD	3
121	14145	14162	18	-	UD	3
273	14151	14162	12	-	HUD	11
126	14151	14167	17	-	UD	3
104	14151	14170	20	-	UD	2
75	14151	14175	25	-	UD	2
273	14157	14168	12	-	HUD	11
320	14159	14170	12	-	HUD	3
260	14177	14189	13	-	UD	2
122	14177	14194	18	-	UD	2
197	14181	14194	14	-	HUD	7
43	14181	14217	37	-	HD	2
280	14187	41498	12	-	HUD	3
341*	14211	14222	12	+	HD	2
203	14227	14240	14	-	UD	3
82	14227	14250	24	-	HD	2
301	14229	14240	12	-	HUD	5
304	14242	14253	12	-	HD	3
139	14242	14258	17	-	UD	2
204	14243	14256	14	-	UD	4
276	14247	14258	12	-	D	2
289	14248	14259	12	-	HD	2
89	14248	14270	23	-	UD	2
286	14253	14264	12	-	HUD	7
165	14253	14267	15	-	UD	5
114	14253	14270	18	-	UD	2
96	14254	14275	22	-	UD	3
218	14261	14273	13	-	UD	3
76	14276	14300	25	-	UD	2
308	14305	14316	12	-	HD	3
313	14313	14324	12	-	HUD	4
340*	14337	14350	14	+	HD	2
314	14396	14407	12	-	UD	2
307	14415	14426	12	-	UD	3
207	14415	14428	14	-	UD	3
156	14415	14430	16	-	UD	2
211	14417	14430	14	-	UD	2
231	14418	14430	13	-	UD	2
31	14418	14462	45	-	UD	2
7	14418	14547	130	-	HD	2
273	14419	14430	12	-	HUD	11
312	14425	14436	12	-	HUD	4
201	14425	14438	14	-	HUD	4
124	14425	14441	17	-	HUD	4
286	14430	14441	12	-	HUD	7
306	14433	14444	12	-	HUD	4
52	14433	14467	35	-	HD	3
66	14435	14462	28	-	HUD	4

Repeat #	HSRL start	HSRL end	Length (bp)	Strand	Subtype	Frequency
19	14440	14513	74	-	HUD	3
50	14479	14513	35	-	HUD	3
35	14479	14520	42	-	HUD	4
108	14481	14499	19	-	HUD	7
159	14515	14530	16	-	HUD	3
327	14522	14533	12	-	HUD	5
88	14522	14544	23	-	HD	3
70	14522	14547	26	-	HUD	3
93	14523	14544	22	-	HUD	4
275	14532	14543	12	-	UD	5
304	14536	14547	12	-	HD	3
323	14544	14555	12	-	UD	2
316	14546	14557	12	-	UD	2
99	14549	14568	20	-	UD	2
25	14550	14610	61	-	HD	2
185	14560	14574	15	-	HUD	3
220	14633	14645	13	-	D	2
72	14633	14657	25	-	UD	3
133	14637	14653	17	-	HUD	4
245	14641	14653	13	-	UD	5
157	14642	14657	16	-	HD	2
120	14685	14702	18	-	UD	3
251	14687	14699	13	-	HUD	4
46	14709	14744	36	-	D	2
69	14719	14744	26	-	HUD	5
56	14719	14752	34	-	HD	2
26	14719	14777	59	-	UD	2
161	14722	14737	16	-	HUD	6
100	14722	14741	20	-	HUD	6
95	14722	14743	22	-	HUD	7
77	14722	14745	24	-	HUD	4
4	14754	14919	166	-	HD	2

[#] denotes hairpin loop in repeat sequence
[*] signifies direct repeats that do not include U, H, D or UD repeats.

APPENDIX 4

DNA / Protein sequence alignments performed in this study.

Appendices 4a – 4c are alignments of DNA sequence from three regions of the *hsr* gene described in section 3.2.4. The DNA sequence of the reference strain 4298 is underlined. S and E denote the start and end of each strain sequence, respectively. Blank spaces represent nucleotides identical to the reference strain 4298, whilst gaps in the sequence alignment are displayed as ~. HSRL coordinates are shown above the alignments in brackets.

Appendix 4a. DNA sequence alignment of the repeat region of nine different *H. mustelae* isolates.

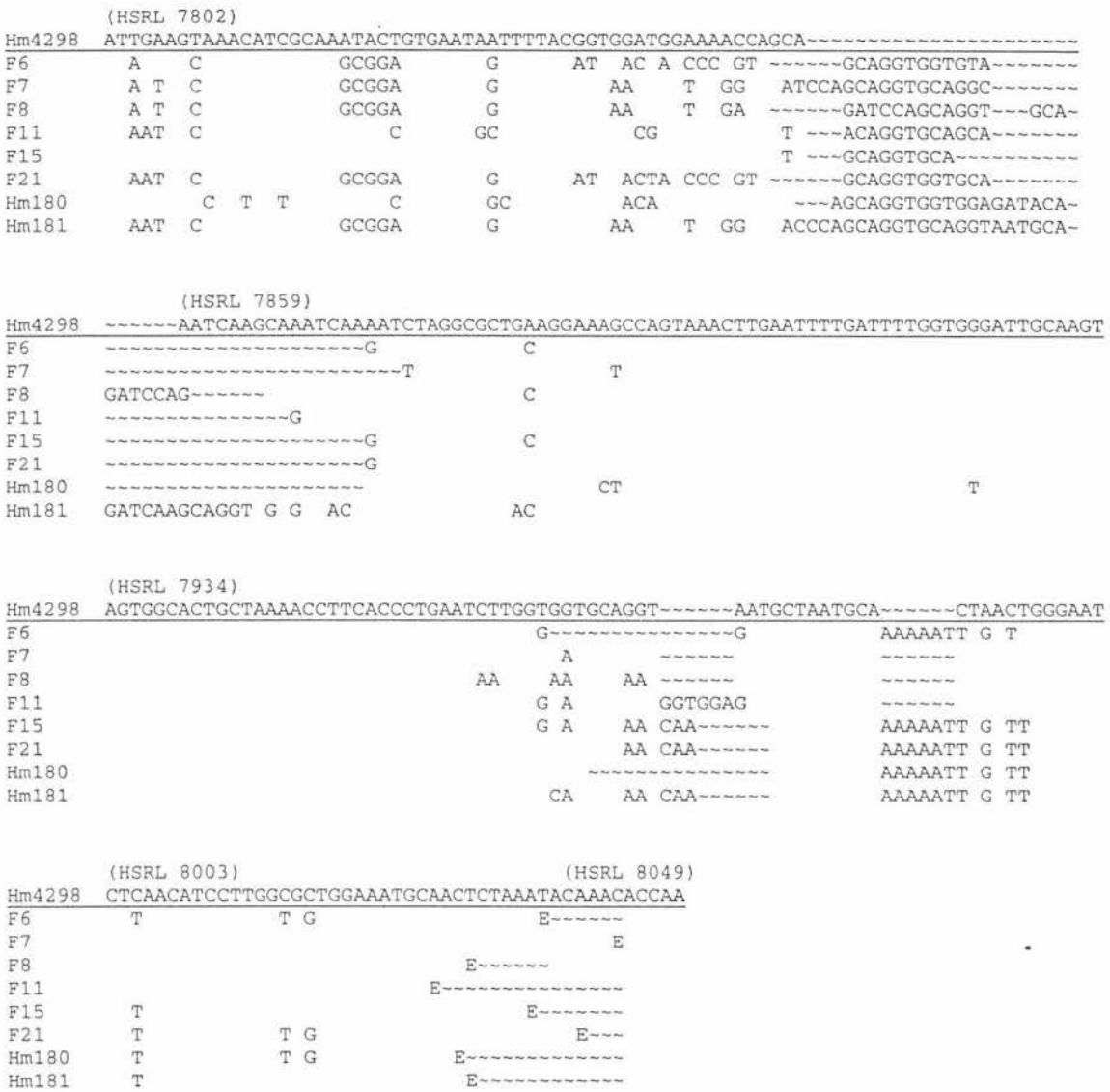
```
          (HSRL 7579)
Hm4298    ~TACTACTCTTGCATTTTCCCTTGCTTCAAGCTTTGTAAATGCAGCAGATGCAGGTAATGCAGGTCAAGCCCCAGTAAAT
F6        ~~~~~~S                                     ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
F7        ~~~~~~~~~~~~S                               ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
F8        ~~~~~~S                                            ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
F11       ~~~~S                                       ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
F15       ~~S                                                             A     A~~~ G
F21       ~~~~~~~~~~~~S                               ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Hm180     ~~~~~S                                             ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
Hm181     ~~~~~~~~~~~~S                                    A~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
```

```
          (HSRL 7658)
Hm4298    GCA~~~GAA~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~GGAATCACAGTGACAGTAAAT~~~CAA
F6        ~~~~~~   ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~              A      ~~~
F7        ~~~~~~   ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~                     ~~~
F8        ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~AC                   ~~~
F11       ~~~~~~   ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~                    AAAT~~~
F15          GATC  GCAGATCAAGCACAACCAGCAGCCCCAGCGCCCGCTGTGGTAAGT                  G AAAT~~~
F21       ~~~~~~   ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~                   CAGCT~~~
Hm180     ~~~~~~C  GCAGATCAAGCACAACAAGCAGCCCCAGCGCCCGCTGTGGTAAGT                    AAAT~~~
Hm181     ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~ C                   ~~~
```

```
          (HSRL 7688)
Hm4298    GCA~~~~~~~~~AAT~~~AAAACTGCGACGGTAAGCGGAAATAACGGGAATGCTACTTTTACATTTACCAAT~~~GGT
F6           GCAAATCCAG  ~~~~~~G       A                                          ~~~
F7           CAAGCAGCA   CTAGGTG    TGAA    AG    GCGG                            ~~~
F8           CAAGCTACA   ~~~C   A      A     T        T                           ~~~
F11          GCAGGTGCA   ~~~C   A   TGAA    AG    GCGG                            ~~~
F15          GCAGATCCA   ~~~C          A     T                                    ~~~
F21          GCAGCTCAA   ACAC       TGAA    AG    GCGG                            ~~~
Hm180        GCAGATGGA   ~~~C                                                     GCA
Hm181        GCAGGTGCA   ~~~C       TGAA          GCGG                            ~~~
```

```
          (HSRL 7751)
Hm4298    GCTAACACCACCGTAAAT~~~GGA~~~~ACAGCAGAC~~~CCA~~~~~~~~~~~~~~~GCAGTA~~~~~~ACAGCTCCAAAC
F6           GG  GTCTGA    AGAT   ~~~~       TA~~~~~~~~~CCAGCAGCAAAT~~~CA AATGCC    GGA G
F7                    A     ~~~   AAT~C     C AGCAAAT~~~~~~~~~~~~~~~C     ~~~~~~    GGA G
F8           GG  GTCTGA GC AGAT   ~~~~       T ACAAAT~~~CAAGCAGCAAATC   A ACTGCC    GGA G
F11                         ~~~   ~~~~       TGACAAAT~~~CAAGCAGCAGATAG AA ACTGCC    GGG G
F15                         ~~~   AAT~C  ~~~~~~GCAAAT~~~CCAGCAGCT~~~C     ~~~~~~     A A
F21          GG  GTCTGA GC AGAT   ~~~~       TA~~~~~~~~~ACACCAGCAAATC  CA ACTGCC    GGA G
Hm180       AGG  GTCTGA    AGAT   ~ACTC     ~~~~~~~~~~~~GCAGGTGCAGCAC  CA AATGCT    GG A
Hm181        GG  GTCTGA GC AGAT   ~~~~       TA~~~~~~~~~CCAGCAGCAAATC  CA ACTGCC    GGA G
```

Appendix 4a. DNA sequence alignment of the repeat region of nine different *H. mustelae* isolates (continued...).



Appendix 4b. DNA sequence alignment of part of the central domain of the hsr gene of nine *H. mustelae* isolates. Only strain-to-strain 4298 differences are shown in the alignment. The abbreviation NZ refers to the entire set of strains isolated from New Zealand ferrets (F6, F7, F8, F11, F15, and F21). "Rest" denotes all strains except 4298.

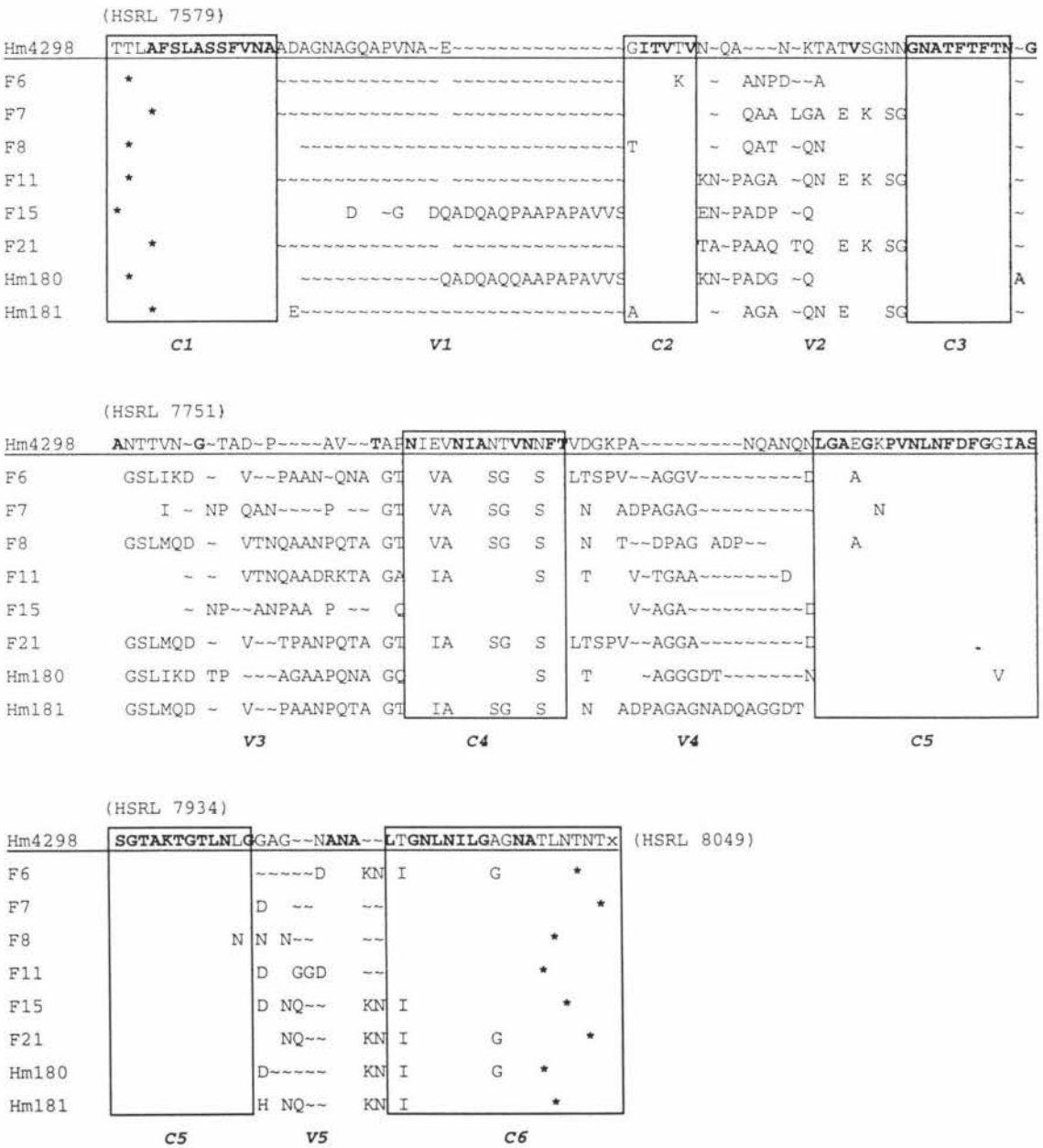
```
      (HSRL 11015)
Hm4298      TATCGATGTT TGCTAGGTAG AGATAGTAGT TGCTAGCAAG TGCTGTTGCT GTTGCGAGTT GGTTGGCCTC
F6                                            S
F7                                          S
F8                                        S
F11                                       S
F15                                         S
F21                                   S
Hm180                                       S
Hm181                                   S

Hm4298      TGAAGTATTT GCCACTGCTT GACTGATGAA GTAAGTAGTA TAGTCTTGAG CATTTGCCTG ATCTTGGCTA
NZ                                                                                    G

Hm4298      CCACCAGTTC CATTGCCACT GCTTGTGCCA CCAGTCCCAC CTAGTCCTGG GATACTACCT AGGCCTGGCG
F6/ F7                                                                          T
F8/ F15                       T         ACA                                     T
F11                     TG A                                                    T
F21                                   C         AG                              T

Hm4298      CTGGAGTACT ATTTCCTGCG TTATTAGCAT TATTTGTTTC TACCTTACCA AAGGCATTTG TTTTTACCGC
Rest                                                              T

Hm4298      AGTAAGTTTT GCATCAAACA CATCAAAACC AATGACAGAA CCCACAGTTG TGAAATTCAC ACTCGCCTGC
F7                                                                                  A
Hm181                                   E

Hm4298      CCACCTTCAT TTTTTACAGT GGCTACCGCA ACGTTACCTG CCCTCTCAGT ACCTCCGCCA TGGTATCTCA
F6/ F15/                                                                          T
F21/ Hm181                                                                        T
F8                                    E                                           T
F11                                                                               E
Hm180                                 E                                           T

Hm4298      CACCAT~AAG ATTTCCATTG CCTAGGATTT GGAGATATTC TGTCACAGAA GTGTTGTTCC TGGGTTTGGT
F6/ F11           C                                             G           ~
F7                C                                             G     C
F15/F21           C                                             G           ~
Hm180/ Hm181      C                                             G           ~

Hm4298      ACACAATCAC ACGGTCACTA TAGATGTTTC CATAGAGTCC AGATCCGTTG AAAGAATTCT GACCATTTAG
F15                                                                               E

Hm4298      AGTCGCGTTC CCCCGCCCAC CCCCAGCACC ATTTCCTTGG TTAGCATCTG CGTTTACATA GACTTTAAAG
F6                        A

Hm4298      AGCGCATTGT TCCCACCAAG TCCTGAAGCT TGTTGCCCAT (HSRL 10345)
F6                E
F7                      E
F21                     ~  G                         E
```

Appendix 4c. DNA sequence alignment of part of the β domain of the *hsr* gene of nine *H. mustelae* strains. Only strain-to-strain 4298 differences are shown in the alignment. "Rest" refers to all strains except strain 4298.

```
     (HSRL 12069)
Hm4298      TGAATCACTA TGAGTAATTT TTTGGAAAAT CACTCAATTT CTTCTGTCTC TTCTTTGTTT TCTTTCTTAT
Rest                       G

Hm4298      CATCTTTTTT GATAGTGTCT TTCATACTAC CTGCAAGAAG AGAAACGTAT TTTCTGTTTT CGCCAAAGTT

Hm4298      ATAGCGATAG CCAATGTTGA ATTGGTAATC TGTGATGATT TTGCCTCCAA AGCTTCTTTC AAAATCAAAG

Hm4298      TAGAATCTAT GATTATCTTT GACGACGAAG TTTGTACCAA TGTTAAAGGC AAATCTACCA GTAGAGCTCA

Hm4298      AAGTCCTTAA AGTGTTTACA GAACCTAGGT CAGACACGAG GGTAAGATTG CCACCACTGA TATAATCGCC
F6                                                                  T
F7                    G                                             T
F15                   G                                             T
F21                                                                 T

Hm4298      GATGTACCAC AAGCCCAAAT AAAGCTCTGA AGCCCAGCCC TTGTCTTCAG TGAATTGATT AAATCTATAA
F6                                                              C
F11                                                             C
F15                                                             C
F21                                                             C

Consensus   CCAAAGTTGC TTCCAATTCG CCCCTGCACT GTGAAGATAG AACTTTGCTC GCCTTTTAGC CAGTGGCTTC

Consensus   CTAGGGTTTG CTTGATATTG CTTTGGTTGA AATAACCTAA AGCAACTTGC CCTTGTGGAG TGATATACCA

Consensus   CTCGTTGTGA TTTCCTAGCA AGAAGCGATA ACCAATCTCT TGAGAGAA (HSRL 11461)
F8                          A
```

Appendix 4d. Amino acid sequence alignment of the repeat region of nine *H. mustelae* isolates.

Strain name is annotated to the left of aligned amino acid sequence. * denotes the start and end of sequence for each strain sequence. Reference strain 4298 is underlined and conserved amino acids are shown in bold. Amino acids identical to reference strain 4298 are omitted whilst gaps in the alignment are denoted with an ~. The corresponding HSRL co-ordinates are shown for strain 4298 above the alignment line. Conserved (C) and variable (V) clusters of residues are grouped together in boxes and labelled beneath in bold italics.



Appendix 4e. Alignment of the DNA sequence derived from the ferret isolate F6 (~800 bp) fragment with the corresponding sequence in the upstream of the *hsr* gene of strain 4298.

The DNA sequence of the reference strain 4298 is underlined. S and E denote the start and end of each strain sequence, respectively. Blank spaces represent nucleotides identical to the reference strain 4298, while gaps in the sequence alignment are displayed as ~. Strain 4298 HSRL coordinates are shown above the alignment.

```
                990       1000      1010      1020      1030      1040      1050
          ....|....|....|....|....|....|....|....|....|....|....|....|....|
Hm4298    GCGATGTTTGTATTTAGAGTTGCATTTCCAGCGCCAAGGATGTTGAGATTCCCAGTTAGTGCATTAGCAT
F6                      S               CCA           A         A C AATTT  T

                1060      1070      1080      1090      1100      1110      1120
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Hm4298    TAGCATTTCCACCACCAAGATTCAGGGTGAAGGTTTTGGCTGGATTGCCAGCAATCCCACCAAAATCAAA
F6              ~~~~~~

                1130      1140      1150      1160      1170      1180      1190
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Hm4298    ATTCAAGTTTACTGGCTTTCCTGTAGCGCCTAGATTTTGATTTGCTGGATCTGCTGGGTCTGCACCTTGT
F6                              TC

                1200      1210      1220      1230      1240      1250      1260
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Hm4298    GCATTTCCACCTGCACCTACTAGCTTGGCATCCCCGTTGCATTTGCGCTATCGTTATTAAATTTTTGCAA
F6

                1270      1280      1290      1300      1310      1320      1330
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Hm4298    ATAATTTTCTATCTGCTGGCCTCTTGCTTGAGCATTAGTATTTCCACTTACACCACCAAGCCCACCTGCT
F6

                1340      1350      1360      1370      1380      1390      1400
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Hm4298    TGTTCCATCAACTTTCCTAGATTGCCAGCAAGCCCAGTAGTAAGCCC~AATTGAAAAATATTTTTGGATT
F6              T                                 T   C       C

                1410      1420      1430      1440      1450      1460      1470
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Hm4298    TGTCGCTGTGATATTTGCAGCGGGCGCTGGGGCTGCTTGATCTGCACCTGCTACTGGATTTCCTTCAAGC
F6                                            G

                1480      1490      1500      1510      1520      1530      1540
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Hm4298    TTGATTTGTGTCTGCACCAGCTGGGTTTGTGCTAGGGGGGCTTTTGTCTGTGTTATCCAATGGATTGCCA
F6

                1550      1560      1570      1580      1590      1600      1610
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Hm4298    TTAAATTTATCCAAATAACCTTGGATCTGCTGGGATTTTGCAATAGCATTAGCACCTGCACCACCTGCA~
F6                                                                          C

                1620      1630      1640      1650      1660      1670      1680
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Hm4298    ~TTCAGCCCGCCTGCTTGATCCATCAAATTTTTTAGATTTGTAGGAGTCCAGTACAAGCCCAATTGAAAA
F6        CAC           E
```

REFERENCES

Alm, R. A., Ling, L. L., Moir, D. T., King, B. L., Brown, E. D., Doig, P. C., Smith, D. R., Noonan, B., Guild, B. C., deJonge, B. L., Carmel, G., Tummino, P. J., Caruso, A., Uria-Nickelsen, M., Mills, D. M., Ives, C., Gibson, R., Merberg, D., Mills, S. D., Jiang, Q., Taylor, D. E., Vovis, G F., and Trust, T. J. (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**:176-80.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* **215**(3):403-10.

Atherton, J. C., Cao, P., Peek, R. M., Tummuru, M. K. R., Blaser, M. J., and Cover, T. L. (1995). Mosaicism in vacuolating cytotoxin alleles of *Helicobacter pylori*. *J Biol Chem* **270**(30):17771-77.

Berghammer, H., and Auer, B. (1993). "Easypreps": fast and easy plasmid minipreparation for analysis of recombinant clones in *E. coli*. *Biotechniques* **14**(4): 524, 528.

Birnboim, H. C., and Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Res* **7**(6):1513-23.

Blaser, M. J. (1993). *Helicobacter pylori*: microbiology of a 'slow' bacterial infection. *Trends Microbiol* **1**(7):255-60.

Blaser, M. J. (1996). The bacteria behind ulcers. *Scientific American* **2**:92-97.

Blomfield, I. C., Kulasekara, D. H., and Eisenstein, B. I. (1997). Integration host factor stimulates both FimB- and FimE-mediated site-specific DNA inversion that controls phase variation of type 1 fimbriae expression in *Escherichia coli*. *Mol Microbiol* **23**:705-17.

Boot, H. J., and Pouwels, P. H. (1996). Expression, secretion and antigenic variation of bacterial S-layer proteins. *Mol Microbiol* **31**(6):1117-23.

Boot H. J., Kolen, C. P. A. M., and Pouwels, P. H. (1996). Interchange of the active and silent S-layer protein genes of *Lactobacillus acidophilus* by inversion of the chromosomal *slp* segment. *Mol Microbiol* **21**:799-809.

Borst, P., and Greaves, D. R. (1987). Programmed gene rearrangements altering gene expression. *Science* **235**:658-67.

Brunham, R. C., Plummer, F. A., and Stephens, R. S. (1993). Bacterial antigenic variation, host immune response, and pathogen-host coevolution. *Infect Immun* **61**(6):2273-76.

Cave, D. R., Taylor, N., Tuczynske, and C., Fox J. (1986). *Campylobacter*-like organisms (CLO) from man and ferret: towards an animal model of CLO induced disease. *Gastroenterology* **90**(5):1368.

Chung, C. T., Niemela, S. L., and Miller, R. H. (1989). One-step preparation of competent *Escherichia coli*: transformation and storage of bacterial cells in the same solution. *Proc Natl Acad Sci U S A* **86**(7):2172-75.

Conner, C. P., Heithoff, D. M., Julio, S. M., Sinsheimer, R, L., and Mahan, M. J. (1998). Differential patterns of acquired virulence genes distinguish *Salmonella* strains. *Proc Natl Acad Sci U S A.* **95**(8):4641-45.

Corn, P. G., Anders, J., Takala, A. K., Kayhty, H., and Hoseith, S. K. (1993). Genes involved in *Haemophilus influenzae* type b capsule expression are frequently amplified. *J Infect Dis* **167**(2):356-64.

Cover, T. L., and Blaser, M. J. (1992). Purification and characterization of the vacuolating toxin from *Helicobacter pylori*. *J Biol Chem* **267**(15):10570-75.

Das, A., and Xie, Y-H. (2000). The Agrobacterium T-DNA transport pore proteins VirB8, Vir9, and VirB10 interact with one another. *J Bacteriol* **182**(3):758-63.

de Cock, H., Struyve, M., Kleerebezem, M., van der Krift, T., and Tommassen, J. (1997). Role of the carboxy-terminal phenylalanine in the biogenesis of outer membrane PhoE of *Escherichia coli* K-12. *J Mol Biol* **269**(4):473-78.

Doenges, J. L. (1938). Spirochaetes in the gastric glands of *Macacus rhesus* and humans without definite history of related disease. *Proc Soc Exp Med Biol* **38**:536-38.

D'Souza, S. E., Ginsberg, M. H., and Plow, E. F. (1991). Arginyl-glycyl-aspartic acid (RGD): a cell adhesion motif. *Trends Biochem Sci* **16**(7):246-50.

Dunn, B. E., Cohen, H., and Blaser, M. J. (1997). *Helicobacter pylori*. *Clin Microbiol Rev* **10**(4):720-41.

Dworkin, J., and Blaser, J. (1996). Generation of *Campylobacter fetus* S-layer diversity utilizes a single promoter on an invertible DNA segments. *Mol Microbiol* **19**:1241-53.

Dworkin, J., and Blaser, J. (1997a). Nested DNA inversion as a paradigm of programmed gene rearrangement. *Proc Natl Acad Sci USA* **94**:985-90.

Dworkin, J., and Blaser, J. (1997b). Molecular mechanisms of *Campylobacter fetus* surface layer protein expression. *Mol Microbiol* **26**:433-40.

Dworkin, J., Shedd, O. L., and Blaser, M.J. 1997). Nested DNA inversion of *Campylobacter fetus* S-layer genes is *recA* dependent. *J Bacteriol* **179**:7523-29.

Erdman, S. E., Correa, P., Coleman, L. A., Schrenzel, M. D., Li, X., Fox, J. G. (1997). *Helicobacter mustelae*-associated gastric MALT lymphoma in ferrets. *Am J Pathol* **151**(1):273-80.

Estrem, S. T., Gaal, T., Ross, W., and Gourse, R. L. (1998). Identification of an UP element consensus sequence for bacterial promoters. *Proc. Natl. Acad. Sci. USA* **95**:9761–66.

Ferrero, R. L. (1997). Immune responses to mucosal infection: the *Helicobacter pylori* paradigm. *Res Immunol* **148**:91-107.

Forester, N. T., Parton, K., Lumsden, J. S., and O'Toole, P. W. (2000). Isolation of *Helicobacter mustelae* from ferrets in New Zealand. *New Zealand Veterinary Journal* **48**:65-69.

Forester, N, Lumsden, J. S., O'Croinin, T., and O'Toole, P. W. (2001). Sequence and antigenic variability of the *Helicobacter mustelae* surface ring protein Hsr. *Infect Imm* **69**(5):3447-50.

Forman, D. (1993). An international association between *Helicobacter pylori* infection and gastric cancer. *Lancet* **341**:1668.

Fox, J. G., Edrise, B. M., Cabot, E. B., Beaucage, C., Murphy, J. C., and Prostak, K. S. (1986). *Campylobacter*-like organisms isolated from gastric mucosa of ferrets. *Am J Vet Res* **47**(2):236-39.

Fox, J. G., Chilvers, T., Goodwin, C. S., Taylor, N. S., Edmonds, P., Sly, L. I., and Brenner, D. J. (1989). *Campylobacter mustelae*. a new species resulting from the elevation of *Campylobacter pylori subsp. mustelae* to species status. *Int J Syst Bacteriol* **39**:301-303.

Fox, J. F., Correa, P., Taylor, N. S., Lee, A., Otto, G., Murphy, J. C., and Rose, R. (1990). *Helicobacter mustelae*-associated gastritis in humans. *Gastroenterology* **99**:352-61.

Fox, J. G., Otto, G., Murphy, J. C., Taylor, N. S., and Lee, A. (1991). Gastric colonization of the ferret with *Helicobacter* species: natural and experimental infections. *Rev Infect Dis* Suppl **8**:S671-80.

Fox, J. G., Paster, B. J., Delwhirst, F. E., Taylor, N. S., Yan, L. –L., Macuch, P. J., and Chmura, L. M. (1992). *Helicobacter mustelae* isolation from feces of ferrets: evidence to support fecal-oral transmission of a gastric *Helicobacter*. *Infect Immun* **60**(2):606-11.

Fox J. G., DeCross, A., Taylor, N. S., Yan, L-L, Dewhirst, F. E., Paster, B. J., Meade, L., Krakowka, S., Gorham, J., Marshall, B. (1993). Initial isolation and characterization of *Helicobacter mustelae* from the stomach of mink (*Mustela vison*). *Acta Gastro-Enterologica Belgica* **56**(Suppl):97.

Fox, J. G., and Lee, A. (1997). The role of *Helicobacter* species in newly recognized gastrointestinal tract diseases of animals. *Lab Animal Sci* **47**(3):222-55.

Fox, J. G., Dangler, C. A., Sager, W., Borkowski, R., and Gliatto, J. M. (1997). *Heliocobacter mustelae*-associated gastric adenocarcinoma in ferrets (*Mustela putorius furo*). *Vet Pathol* **34**(3):225-29.

Fox, J. G. (1998). *Biology and diseases of the ferret, 2nd edition*. Lippincott, Williams & Wilkins.

Gaal, T., Rao, L., Estrem, S. T., Yang, J., Wartell, R. M., and Gourse, R. L. (1994). Localization of the intrinsically bent DNA region upstream of the *E.coli rrnB* P1 promoter. *Nucleic Acids Res* **22**(12):2344-50.

Gibbs, C. P., Reimann, B., Schultz, E., Kaufmann, A., Haas, R., and Meyer, T. F. (1989). Reassortment of pilin genes in *Neisseria gonorrhoeae* occurs by two distinct mechanisms. *Nature* **338**:651-52.

Gilsdorf, J. R. (1998). Antigenic diversity and gene polymorphisms in *H. influenzae*. *Infect Immun* **66**(11):5053-59.

Goodwin, C. S., McCulloch, R. K., Armstrong, J. A., and Wee, S. H. (1985). Unusual cellular fatty acids and distinctive ultrastructure in a new sprial bacterium (*Campylobacter pyloridis*) from the human gastric mucosa. *J Med Microbiol* **19**:257-67.

Goodwin, C. S., Armstrong, J. A., Chilvers, T., Peters, M., Collins, M. D., Sly L., McConnel, W., and Harper, W.E.S. (1989). Transfer of *Campylobacter pylori* and *Campylobacter mustelae* to *Helicobacter* gen. nov. as *Helicobacter pylori* comb. nov. and *Helicobacter mustelae* comb. nov. respectively. *Int J Sys Bacteriol* **39**(4):297-405.

Goto, K., Ohashi, H., Ebukuro, S., Itoh, K., Tohma, Y., Takakura, A., Wakana, S., Ito, M., and Itoh, T. (1998). Isolation and characterization of *Helicobacter* species from the stomach of the house musk shrew (*Suncus murinus*) with chronic gastritis. *Curr Microbiol* **37**(1):44-51.

Haas, R., and Meyer, T. F. (1986). The repertoire of silent pilus genes in *Neisseria gonorrhoeae*: evidence for gene conversion. *Cell* **44**:107-15.

Haas, R., Schwarz, H., and Meyer, T. F. (1987). Release of soluble pilin antigen coupled with gene conversion in *Neisseria gonorrhoeae*. *Proc Natl Acad Sci USA* **84**:9079-83.

Hacker, J., Blum-Oehler, G., Muhldorfer, I., and Tschape, H. (1997). Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. *Mol Microbiol* **23**(6):1089-97.

Hall, T.A. (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl Acids Symp Ser* **41**:95-98.

Hammar, M., Tyszkiewicz, T., Wadstrom, T., and O'Toole, P.W. (1992). Rapid detection of *Helicobacter pylori* in gastric biopsy material by polymerase chain reaction. *J Clin Microbiol* **30**(1):54-58.

Hamrick, T. S., Dempsey, J. A., Cohen, M. S., and Cannon, J. G. (2001). Antigenic variation of gonococcal pilin expression *in vivo*: analysis of the strain FA1090 pilin repertoire and identification of the *pilS* gene copies recombining with *pilE* during experimental human infection. *Microbiology* **147**(4):839-49.

Hanahan, D. (1983). Studies on transformation of *Escherichia coli* with plasmids. *J Mol Biol* **166**(4):761-69.

Harvey, C. B., and Reynolds, R. P. (1987). Analysis of *E. coli* promoter sequences. *Nucleic Acids Res* **15**(5):2343-61.

Henderson, I. R., Meehan, M., and Owen, P. (1997). A novel regulatory mechanism for a novel phase-variable outer membrane protein of *Escherichia coli*. *Adv Exp Med Biol* **49**:349-55.

Henderson, I. R., Navarro-Garcia, F., and Nataro, J. P. (1998). The great escape: structure and function of the autotransporter proteins. *Trends Microbiol* **6**(9):370-78.

Henderson, I. R., Owen, P., and Nataro, J. P. (1999). Molecular switches – the ON and OFF of bacterial phase variation. *Mol Microbiol* **33**(5), 919-32.

Hendrixson, D. R., de la Morena, M. L., Stathopoulos, C., St Geme, and J. W. 3rd. (1997). Structural determinants of processing and secretion of the *Haemophilus influenzae hap* protein. *Mol Microbiol* **26**(3):505-18.

Hessey, S. J., Spencer, J., Wyatt J.I., Sobala, G., Rathbone, B. J., Axon, A. T., and Dixon, M. F. (1990). Bacterial adhesion and disease activity in *Helicobacter*-associated chronic gastritis. *Gut* **31**:134-38.

Higgins, D. G., Bleasby, A. J., and Fuchs R. (1992). CLUSTAL V: improved software for multiple sequence alignment. *CABIOS* **8**(2):189-91.

Hodges-Garcia, Y., and Hagerman, P. J. (1992). Cytosine methylation can induce local distortions in the structure of duplex DNA. *Biochemistry* **31**(33):7595-99.

Hofreuter, D., Odenbreit, S., Henke, G., and Haas, R. (1998). Natural competence for DNA transformation in *Helicobacter pylori* – identification and genetic characterization of the *comB* locus. *Mol Microbiol* **28**:1027-38.

Hofreuter, D., Odenbreit, S., Puls, J., Schwan, and D., Haas. R. (2000). Genetic competence in *Helicobacter pylori*: mechanisms and biological implications. *Res Microbiol* **151**(6):487-91.

Hoseith, S. K., Moxon, E. R., and Silver, R. P. (1986). Genes involved in *Haemophilus influenzae* type b capsule expression are part of an 18-kilobase tandem duplication. *Proc Natl Acad Sci USA* **83**:1106-10.

Howell-Adams, B., and Seifert, H. S. (1999). Insertion mutations in *pilE* differentially alter gonococcal pilin antigenic variation. *J Bacteriol* **181**(19):6133-41.

Howell-Adams, B., and Seifert, H. S. (2000). Molecular models accounting for the gene conversion reactions mediating gonoccoal pilin antigenic variation. *Mol Microbiol* **37**(5):1146-58.

Hueck, C. J. (1998). Type III protein secretion systems in bacterial pathogens of animals and plants. *Microbiol Mol Biol Rev* **62**(2): 379-433.

Jones, D. M., Curry, A., and Fox, A. J. (1985). An ultrastructural study of the gastric *Campylobacter*-like organism "*Campylobacter pyloridis*". *J Gen Microbiol* **19**:2335-41.

Jose, J., Kramer, J., Klauser, T., Pohlner, J., and Meyer, T. F. (1996). Absence of periplasmic DsbA oxidoreductase facilitates export of cysteine-containing passenger proteins to the *Escherichia coli* cell surface via the Iga beta autotransporter pathway. *Gene* **178**:107-10.

Kersulyte, D., Chalkayskas, H., and Berg, D. E. (1999). Emergence of recombinant strains of *Helicobacter pylori* during human infection. *Mol Microbiol* **31**:31-43.

Klauser, T., Kramer, J., Otzelberger, K., Pohlner, J., and Meyer, T. F. (1993). Characterization of the *Neisseria* Iga beta-core. The essential unit for outer membrane targeting and extracellular protein secretion. *J Mol Biol* **234**:579-93.

Komano, T. (1999). Shufflons: Multiple inversion systems and integrons. *Ann Rev Genet* **33**:171-91.

Koster, M., Bitter, W., de Cock, H., Allaoui, A., Cornelis, G. R., and Tommassen, J. (1997). The outer membrane component, YscC, of the Yop secretion machinery of Yersinia enterocolititica forms a ring-shaped multimeric complex. *Mol Microbiol* **26**(4):789-97.

Kumar, A., Malloch, R. A., Fujita, N., Smillie, D. A., Ishihama, A., and Hayward, R. S. (1993). The minus 35 recognition region of *Escherichia coli* sigma 70 is inessential for initiation of transcription at an 'extended minus 10' promoter. *J Mol Biol* **232**:406-18.

Kyte, J., and Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**:105-32.

Labigne-Roussel, A., Courcoux, P., and Tompkins, L. (1988). Gene disruption and replacement as a feasible approach for mutagenesis of *Campylobacter jejuni*. *J Bacteriol* **170**:1704-708.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**(259):680-85.

Lee, A, Fox, J. G., and Hazell, S. (1993). Pathogenicity of *Helicobacter pylori* – a perspective. *Infect Immun* **61**: 1601-10.

Lee A. (2000). Animal models of gastroduodenal ulcer disease. *Baillieres Best Pract Res Clin Gastroenterol* **14**(1):75-96.

Loveless B. J., and Saier, M. H. Jnr. (1997). A novel family of channel-forming autotransporting, bacterial virulence factors. *Mol Memb Biol* **14**:113-23.

Lupski, J. R., and Weinstock, (1992). Short, interspersed repetitive DNA sequences in Prokaryotic genomes. *J Bacteriol* **174**(14):4525-29.

Marshall, B. J., and Warren, J. R. (1983). Unidentified curved bacilli on gastric epithelium in active chronic gastritis. *Lancet* **1**:1273-75.

Marshall, B. J., and Warren, J. R. (1984). Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* **1**(8390):1311-15.

Marshall B. J. (1995). *Helicobacter pylori*. The etiological agent for peptic ulcer. *JAMA* **274**:1064-66.

Marshall D. G., Dundon, W. G., Beesley, S. M., and Smyth, C. J. (1998). *Helicobacter pylori* – a conundrum of genetic diversity. *Microbiology* **144**:2925-39.

Matsuyama S, Tajima T, and Tokuda H. (1995). A novel periplasmic carrier protein involved in the sorting and transport of Escherichia coli lipoproteins destined for the outer membrane. *EMBO J* **14**(14):3365-72.

Matsuyama, S., Yokota, N., and Tokuda, H. (1997) A novel outer membrane lipoprotein, LolB (HemM), involved in the LolA (p20)-dependent localization of lipoproteins to the outer membrane of *Escherichia coli*. *EMBO J* **16**(23):6947-55.

Maurer, J., Jose, J., and Meyer, T. F. (1999). Characterization of the essential transport function of the AIDA-I autotransporter and evidence supporting structural predictions. *J Bacteriol* **181**(22):7014-20.

McLaren, R. S., Newbury, S. F., Dance, G. S., Causton, H. C., and Higgins, C. F. (1991). mRNA degradation by processive 3' – 5' exoribonucleases *in vitro* and the implications for prokaryotic mRNA decay *in vivo*. *J Mol Biol* **221**(1):81-95.

Mehr, I. J., and Seifert, H. S. (1998). Differential roles of homologous recombination pathways in *Neisseria gonorrhoeae* pilin antigenic variation, DNA transformation, and DNA repair. *Mol Microbiol* **30**(4):697-710.

Melito, P. L., Munro, C., Chipman, P. R., Woodward, D. L., Booth, T. F., and Rodgers, F. G. (2001). *Helicobacter winghamensis* sp. nov., a novel *Helicobacter* sp. isolated from patients with gastroenteritis. *J Clin Microbiol* **39**(7):2412-17.

Meyer, T. F., Gibbs, C. P., and Haas, R. (1990). Variation and control of protein expression in *Neisseria*. *Ann Rev Microbiol* **44**:451-57.

Morgan, D. R., Fox, J. G., and Leunk, R. D. (1991). Comparison of isolates of *Helicobacter pylori* and *Helicobacter mustelae*. *J Clin Microbiol* **29**(2):395-97.

Morris, A., Thomasen, L., Tasman-Jones, C., Nicholson, G., and Heap, M. (1988). Failure to detect gastric *Campylobacter*-like organisms in a group of ferrets in New Zealand. *New Zealand Medical Journal.* **101**:275.

Moses, E. K., Good, R. T., Sinistaj, M., Billington, S. J., Langford, C. J., and Rood, J. I. (1995). A multiple site-specific DNA-inversion model for the control of Omp1 phase and antigenic variation in *Dichelobacter nodosus*. *Mol Microbiol* **17**(1):183-96.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int J Neural Syst* **8**(5-6):581-99.

Newbury, S. F., Smith, N. H., Robinson, E. C., Hiles, I. D., and Higgins, C. F. (1987a). Stabilization of translationally active mRNA by prokaryotic REP sequences. *Cell* **48**:297-310.

Newbury, S. F., Smith, N. H., and Higgins, C. F. (1987b). Differential mRNA stability controls relative gene expression within a polycistronic operon. *Cell* **51**(6):1131-43.

Noormohammadi, A. H., Markham, P. F., Kanci, A., Whithear, K. G., and Browning, G. F. (2000). A novel mechanism for control of antigenic variation in the haemagglutinin gene family of My*coplasma synoviae*. *Mol Microbiol* **35**:911-23.

Odenbreit, S., Till, M., Hofreuter, D., Faller, G., and Haas, R. (1999). Genetic and functional characterization of the *alpAB* gene locus essential for the adhesion of *Helicobacter pylori* to human gastric tissue. *Mol Microbiol* **31**(5):1537-48.

O'Rouke, J., Lee, A., and Fox, J. G. (1992). An ultrastructural study of *Helicobacter mustelae* and evidence of a specific association with gastric mucosa. *J Med Microbiol* **36**:420-27.

O'Toole, P. W., Austin, J. W., and Trust, T. J. (1994). Identification and molecular characterization of a major ring-forming surface protein from the gastric pathogen *Helicobacter mustelae*. *Mol Microbiol* **11**(2):349-61.

O'Toole, P. W., Kostrzynska, M., Trust, T. J. (1994b). Non-motile mutants of *Helicobacter pylori* and *Helicobacter mustelae* defective in flagellar hook production. *Mol Microbiol* **14**(4):691-703.

Otto, G., Fox, J. G., Wu, P. Y., Taylor, N. S. (1990). Eradication *of Helicobacter mustelae* from the ferret stomach: an animal model of *Helicobacter* (*Campylobacter*) *pylori* chemotherapy. *Antimicrob Agents Chemother* **34**(6):1232-36.

Owen, P., Meehan, M., de Loughry-Doherty, H., and Henderson I. (1996). Phase-variable outer membrane proteins in *Escherichia coli*. *FEMS Immunol Med Microbiol* **16**(2):63-76.

Parkhill, J., Wren, B. W., Mungall, K., Ketley, J. M., Churcher, C., Basham, D., Chillingworth, T., Davies, R. M., Feltwell, T., Holroyd, S., Jagels, K., Karlyshev, A. V., Moule, S., Pallen, M. J., Penn, C. W., Quail, M, A., Rajandream, M-A., Rutherford, K. M., van Vilet, A. H. M., Whitehead, S., and Barrell, B. G. (2000a). The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature* **403**:665-68.

Parkhill, J., Achtman, M., James, K. D., Bentley, S. D., Churcher, C., Klee, S. F., Morelli, G., Basham, D., Brown, D., Chillingworth, T., Davies, R. M., Davis, P., Devlin, K., Feltwell T., Hamlin, N., Holroyd, S., Jagels, K., Leather, S., Moule, S., Mungall, K., Quail, M, A., Rajandream, M-A., Rutherford, K. M., Simmonds, M., Skelton, J., Whitehead, S., Spratt, B. G., and Barrell, B. G. (2000b). Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature* **404**:502-506.

Paster, B. J., Lee, A., Albrandt, K. G., Dewhirst, F. E., Tordoff, L. A., Fraser, G. J., O'Rourke, J. L., Taylor, N. S., and Ferrero, R. (1991). The phylogeny of *Helicobacter felis* sp. nov., *Helicobacter mustelae*, and related bacteria. *Int J Syst Bacteriol* **41**:31-38.

Patterson, M. M., Schrenzel, M. D., Feng, Y., Xu, S., Dewhirst, F. E., Paster, B. J., Thibodeau, S. A., Versalovic, J, and Fox, J. G. (2000). *Helicobacter aurati* sp. nov., a urease-positive *Helicobacter* species cultured from gastrointestinal tissue of Syrian hamsters. *J Clin Microbiol* **38**(10):3722-28.

Perkins, S. E., Fox, J. G., Walsh, J. H. (1996). *Helicobacter mustelae*-associated hypergastrinemia in ferrets (*Mustela putorius furo*). *Am J Vet Res* **57**(2):147-50.

Pitcher, D. G., Saunders, N. A., and Owen, R. J. (1989). Rapid extraction of bacterial genomic DNA with guanidium thiocyanate. *Lett Appl Microbiol* **8**:151-56.

Plasterk, R. H. A., Simon, M. I., and Barbour, A. G. (1985). Transposition of structural genes to an expression sequence on a linear plasmid causes antigenic variation in the bacterium *Borrelia hermsii*. *Nature* **318**:257-63.

Pugsley, A. P. (1993). The complete general secretory pathway in Gram-negative bacteria. *Microbiol Rev* **57**(1):50-108.

Romaniuk, P. J., Zoltowska, B., Trust, T. J., Lane, D. J., Olsen, G. J., Pace, N. R. and Stahl, D. A. (1987). *Campylobacter pylori*, the spiral bacterium associated with human gastritis, is not a true *Campylobacter* sp. *J Bacteriol* **169**(5):2137-41.

Ross, W., Gosink, K. K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., Severinov, K., and Gourse, R. L. (1993). A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science* **262**(5138):1407-13.

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989). *Molecular cloning: a laboratory manual. 2nd Ed.*, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Sanger, F., Donelson, J. E., Coulson, A. R., Kossel, H. and Fischer, D. (1974). Determination of a nucleotide sequence in bacteriophage f1 DNA by primed synthesis with DNA polymerase. *J Mol Biol* **90**(2):315-33.

Sara, M., and Sleytr, U. B. (2000). S-layer proteins. *J Bacteriol* **182**(4):859-68.

Sakari, J., Pandit, N., Moxon, E. R., and Achtman, M. (1994). Variable expression of the Opc outer membrane protein in *Neisseria meningitidis* is caused by size variation of a promoter containing poly-cytidine. *Mol Micrbiol* **13**:207-17.

Saunders N. J., Peden J. F., and Moxon E. R. (1999). Absence in *Helicobacter pylori* of an uptake sequence for enhancing uptake of homospecific DNA during transformation. *Microbiology* **145**(12):3523-28.

Seifert, H. S., and So, M. (1988). Genetic mechanisms of bacterial antigenic variation. *Microbiol Rev* **52**(3):327-36.

Seydel, A., Gounon, P., and Pugsley, A.P. (1999). Testing the '+2 rule' for lipoprotein sorting in the *Escherichia coli* cell envelope with a new genetic selection. *Mol Microbiol* **34**(4):810-21.

Shannon, J. L., and Fernandez, R. C. (1999). The C-terminal domain of the *Bordetella pertussis* autotransporter BrkA forms a pore in lipid bilayer membranes. *J Bacteriol* **181**(18):5838-42.

Smeets, L. C., Bijlsma, J. J. E., Kuipers, E. J., Vandenbroucke-Grauls, C. M. J. E.,and Kusters J. G. (2000). The *dprA* gene is required for natural transformation of *Helicobacter pylori*. *FEMS Immunol Med Microbiol* **27**:99–102.

Smeets, L. C., Bijlsma, J. J., Boomkens, S. Y., Vandenbroucke-Grauls, C. M., and Kusters, J. G. (2001). *comH*, a novel gene essential for natural transformation of *Helicobacter pylori*. *J Bacteriol* **182**(14):3948-54.

Sokurenko, E. V., Hasty, D. L., and Dykhuizen D. E. (1999). Pathoadaptive mutations: gene loss and variation in bacterial pathogens. *Trends Microbiol* **7**(5):191-95.

Southern, E. M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol* **98**(3):503-17.

Steer, H. W., and Colin-Jones, D. G. (1975). Mucosal changes in gastric ulceration and their response to carbenoxolone sodium. *Gut* **16**:590-97.

Stern, M. J., Ames, G. F-L., Smith, N. H., Robinson, E. C., and Higgins, C. F. (1984). Repetitive extragenic palindromic sequences: a major component of the bacterial genome. *Cell* **37**:1015-26.

Suzuki, T., Lett, M. C., and Sasakawa, C. (1995). Extracellular transport of VirG protein in shigella. *J Biol Chem* **270**(52):30874-80.

Tajima T., Yokota, N., Matsuyama, S., and Tokuda, H. (1998). Genetic analyses of the in vivo function of LolA, a periplasmic chaperone involved in the outer membrane localization of *Escherichia coli* lipoproteins. *FEBS Lett* **439**(1-2):51-54.

Tettelin, H., Saunders, N. J., Heidelberg, J., Jeffries, A. C., Nelson, K. E., Eisen, J. A., Ketchum, K. A., Hood, D. W., Peden, J. F., Dodson, R. J., Nelson, W. C., Gwinn, M. L., DeBoy, R., Peterson, J. D., Hickey, E. K., Haft, D. H., Salzberg, S. L., White, O., Fleischmann, R. D., Dougherty, B. A., Mason, T., Ciecko, A., Parksey, D. S., Blair, E., Cittone, H., Clark, E. B., Cotton, M. D., Utterback, T. R., Khouri, H., Qin, H., Vamathevan, J., Gill, J., Scarlato, C., Masignani, V., Pizza, M., Grandi, G., Sun, L., Smith, H. O., Fraser, C. M., Moxon, E. R., Rappuoli, R., and Venter, J. C. Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. (2000). *Science* **287**:1809-15.

Therisod, H., Monteiro, M. A., Perry, M. B., and Caroff, M. (2001). *Helicobacter mustelae* lipid A structure differs from that of *Helicobacter pylori*. *FEBS Lett* **499**:1-5.

Thomas, S. R., and Trust, T. J. (1995). Tyrosine phosphorylation of the tetragonal paracrystalline array of *Aeromonas hydrophila*: Molecular cloning and high-level expression of the S-layer protein gene. *J Mol Biol* **245**:568-81.

Tomb, J-F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., Fitzegerald, L. S., Lee, N., Adams, M. D., Hickey, E. K., Berg, D. E., Gocayne, J. D., Utterback, T. R., Peterson, J. D., Kelley, J. M., Cotton, M, D., Weldman, J. M., Fujii C., Bowman, C., Watthey, L., Wallin, E., Hayes, W. S.,

Borodovsky, M., Karp, P. D., Smith, H. O., Fraser, C. M., and Venter, J. C. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**:539-47.

Tompkins, D. S., Wyatt, J. I., Rathbone, B. J., and West, A. P. (1988). The characterization and pathological significance of gastric *Campylobacter*-like organisms in the ferret: a model for chronic gastritis? *Epidemiol Infect* **101**(2):269-78.

Towbin, H., Staehelin, T., and Gordon, J. (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: Procedure and some applications. *Proc Natl Acad Sci U S A* **76**:4350-54.

van Belkum, A., Scherer, S., van Alphen, L., and Verbrugh, H. (1997). Short-sequence repeats in prokaryotic genomes. *Microbiol Mol Biol Rev* **62**(2):275-93.

van Belkum, A., Scherer, S., van Leeuwen, W., Willemse, D., van Alphen, L., and Verbrugh, H. (1998). Variable number of tandem repeats in clinical strains of *Haemophilus influenzae*. *Infect Immun* **65**(12):5017-27.

van der Ende, A., Hopman, C. T. P., Zaat, S., Oude Essink, B. B., Berkhout, B., and Dankert, J. (1995). Variable expression of class 1 outer membrane protein in *Neisseria meningitidis* is caused by variation in the spacing between the –10 and –35 regions of the promoter. *J Bacteriol* **177**(9):2475-80.

Veiga, E., de Lorenzo, V., and Fernandez, L. A. (1999). Probing secretion and translocation of a beta-autotransporter using a reporter single-chain Fv as a cognate passenger domain. *Mol Microbiol* **33**(6):1232-43.

Weiser, J. N., Maskell, D. J., Butler, P. D., Lindberg, A. A., and E. R., Moxon. (1990). Characterization of repetitive sequences controlling phase variation of *Haemophilus influenzae* lipopolysaccharide. *J Bacteriol* **172**:3304-309.

Wren B. W. (2000). Microbial genome analysis: insights into virulence, host adaptation and evolution. *Nat Rev Genet* **1**(1):30-39.

Yakushi, T., Yokota, N., Matsuyama, and S., Tokuda, H. (1998). LolA-dependent release of a lipid-modified protein in the inner membrane of *Escherichia coli* requires nucleoside triphosphate. *J Biol Chem* **273**(49):32576-81.

Yakushi, T., Masuda, K., Narita, S., Matsuyama, S., and Tokuda, H. (2000). A new ABC transporter mediating the detachment of lipid-modified proteins from membranes. *Nat Cell Biol* **2**(4):212-18.

Yanisch-Perron, C., Vieira, J., and Messing J. (1985). Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* **33**(1):103-19.

Yasuno, K., Yamazaki, T., Tanaka, Y., Kodama, T. S., Matsugami, A., Katahira, M., Ishihama, A., and Kyogoku, Y. (2001). Interaction of the C-terminal domain

of the *E. coli* RNA polymerase alpha subunit with the UP element: recognizing the backbone structure in the minor groove surface. *J Mol Biol* **306**(2):213-25.

Yokota, N., Kuroda, T., Matsuyama, S., and Tokuda, H. (1999). Characterization of the LolA-LolB system as the general lipoprotein localization mechanism of *Escherichia coli*. *J Biol Chem* **274**(43):30995.

Zhang, Q. Y., DeRyckere, D., Lauer, P., and Koomey, M. (1992). Gene conversion in *Neisseria gonorrhoeae*: Evidence for its role in pilus antigenic variation. *Proc Natl Acad Sci USA* **89**:5366-70.