Variant-specific surface protein

(VSP) gene subsets in *Giardia*

by

Mandana Mansouri B.Sc. (Adelaide)

A thesis submitted for the degree of Doctor of Philosophy

Department of Molecular Biosciences
The University of Adelaide

December 2000
(amended May 2001)

Table of Contents

Declaration

This thesis contains no material which has been accepted for the award of any other degree or diploma in any University or other tertiary institution and to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference is made in the text.

I give consent for this copy of my thesis, when deposited in the University library, to be available for loan and photocopying.

Mandana Mansouri

December 9, 2000

Acknowledgments

I would like to thank my supervisors Dr. Peter Ey and Assoc. Prof. Graham Mayrhofer for giving me the opportunity to undertake my postgraduate studies. Especially I would like to thank Peter for his guidance, encouragement and help with this thesis.

I am grateful to Jocelyn Darby for her technical assistance in the lab, particularly with *in situ* hybridisation.

Thanks to Gail, Shelley, Ros, John, Gary, Chris and Tony for their help in many different areas and for the friendly conversations.

Further thanks to the students and staff of the Department of Molecular Biosciences, Bec, Gorjana, Melissa, Xingqi, Craig, Angela, David and Maria for their friendship throughout my postgraduate study.

I would also like to thank my Mum and Dad, who have been always there for me. I will never forget all that you have done and your continuous encouragement.

Finally, and most importantly, to my husband, Mehdi, a very special thank you for your love, support, patience and help in looking after Ramin.

To Ramin, my beautiful son, thank you for your patience in the past few months. I can now play with you every night as long as you like.

Abstract

This project was concerned with the family of genes that encode variant-specific surface proteins (VSP) in the parasitic protozoon, *Giardia*. The VSP are immunodominant surface antigens on *Giardia* trophozoites and the alternative expression of different *vsp* genes underlies observations of antigenic variation within *Giardia* populations. Specific goals were to identify and compare closely related genes, which may form subsets within the larger VSP gene family. Fewer than 10 of a possible repertoire of 150-300 VSP genes had been characterised completely at the time the project was commenced.

Two novel VSP gene subfamilies were identified in the study. One, designated the *vsp*136 subfamily (for which the previously described gene, *crp*136, is the prototype), comprises genes with tandem repeat elements. The other, designated the *vsp*72 subfamily, has members that lack tandem repeat elements and includes numerous pseudogenes. Most of the genes examined were identified among 21 distinct plasmids isolated by screening *Giardia* genomic DNA libraries using hybridisation analysis. Eight genomic DNA fragments (size range, 3.5-9.5 kb) were characterised completely by nucleotide sequence determinations. Another 15 fragments (3.5-15 kb) were characterised partially by a combination of restriction fragment length polymorphism (RFLP), Southern hybridisation and polymerase chain reaction (PCR) analysis. Amplified segments (1-1.6 kb, derived from both genomic and complementary DNA) from an additional 7 *vsp* gene loci were also characterised.

Analysis of individual and aligned nucleotide sequences and the polypeptides (functional and pseudo-VSP) encoded by these various genes revealed conserved segments that allowed the identification of distinct gene subfamilies. Characterised members of the *vsp*136 subfamily contain different copy numbers of related (in some cases, identical) tandem repeat elements. These genes show clear evidence of recombination. The *vsp*72 subfamily consists of tandemly arrayed pseudogenes (of which 13 were identified in this study) as well as functional genes, all of which lack tandem repeat elements. To complement the genetic analysis, mRNA *in situ* hybridisation experiments were performed on *Giardia* trophozoites to identify cells expressing functional genes belonging to (or derived by recombination from nonfunctional members of) these subfamilies. In all, nucleotide sequences comprising a total of 43,568 bp and involving 23 loci (of which 10 were characterised completely) were determined during the course of this project.

List of Abbreviations

Abbreviations used in this thesis are defined below and also at first usage within the text.

µg	microgram(s)
µl	microliter(s)
µm	micrometer(s)
µM	micromolar
°C	degrees Celsius
%	percentage
Amp	ampicillin
bp	base-pair(s)
cds	coding sequence
CRP	cysteine-rich protein(s)
dATP	2'-deoxyadenosine 5'-triphosphate
dCTP	2'-deoxycytidine 5'-triphosphate
dGTP	2'-deoxyguanosine 5'-triphosphate
dTTP	2'-deoxythymidine 5'-triphosphate
dNTP	2'-deoxyribonucleoside 5'-triphosphate(s)
DIG	digoxigenin
DNA	deoxyribonucleic acid
g	gram(s)
hr	hour(s)
kDa	kilodalton(s)
M	molar
mAb	monoclonal antibody /ies
mg	milligram(s)
min	minute(s)
ml	milliliter(s)
mM	millimolar
mRNA	messenger RNA
MW	molecular weight
ng	nanogram(s)

Abbreviations (continued)

nt	nucleotide(s)
Oligo(s)	oligodeoxynucleotide(s)
ORF	open reading frame(s)
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction(s)
Poly(A)	polyadenylation / polyadenylated
RFLP	restriction fragment length polymorphism(s)
RNA	ribonucleic acid
RNase	Ribonuclease
rRNA	ribosomal RNA
RT	reverse transcriptase
SDS	sodium dodecylsulphate
Tris	tris(hydroxymethyl)aminomethane
v; v/v; w/v	volume; volume to volume; weight to volume
VSP	variant-specific surface protein(s)

Chapter 1

Introduction

1.1 Giardiasis

Giardiasis is a gastrointestinal disease caused by the parasitic protozoon *Giardia intestinalis* (syn. *G. lamblia*), considered by the World Health Organisation and by many national governments (e.g. U.S.A, Canada and New Zealand) to be one of the most important intestinal protozoon parasites because of the morbidity caused by the disease and the incidence of infected individuals (Acha & Szyfres, 1987; Thompson *et al.* 1990; 1993; Meyer, 1990; Adam, 1991).

The prevalence of individuals testing positive for *Giardia* cysts in stool samples varies between 2 and 5% in the industrialised world and up to 20-30% in the developing world. In the U.S.A, *Giardia* is the most commonly reported pathogenic protozoon (Kappus *et al.* 1994). Infections are especially common in third-world countries and predominantly in children (Meyer, 1990). *Giardia* infections are so prevalent in rural areas of the Gambia that the peak age for the appearance of *Giardia*-specific IgM antibodies in the plasma of the infants is 3-4 months (Lunn *et al.* 1999). It has been reported in other studies that 40% of Peruvian children become infected with *Giardia* in their first 6 months of life, and that all rural Guatemalian children are infected before the age of 3 (Miotti et al, 1986; Mata, 1978). The transmission of giardiasis occurs by the faecal-oral route, most commonly by direct (person-to-person) contact, e.g. in child-care centres (Thompson, 1994), or less frequently, by venereal transmission (Schemerin *et al.* 1978; Grimmond *et al.* 1988; Steketee *et al.* 1989). High-dose waterborne or food borne transmission seems, at least in western countries where this has been studied, to be less frequent although single contamination events have the potential to cause sudden, small-scale epidemics (Quick *et al.* 1992; Anderson *et al.* 1993; Isaac-Renton *et al.* 1993).

Giardia infections may be acute or chronic and they can be asymptomatic (possibly the majority of infections) or produce clinical disease characterised by steatorrhoea, diarrhoea, nausea, abdominal discomfort and weight loss (Moore *et al.* 1969; Brodsky *et al.*

1974). It is not clear why some individuals infected with *Giardia* remain asymptomatic while others become ill, although the severity and nature of the disease is clearly influenced by multiple factors (Farthing, 1992, 1997; Wolfe, 1992). There is as yet no obvious correlation between symptomatic disease and infections by particular subtypes of *G. intestinalis*, but it remains possible that differences in the development (or absence) of giardiasis is influenced by the particular strain (or strains) of *Giardia* that (co)infect an individual and by the repertoire of expressed surface antigens in addition to obvious host factors such as age, immunocompetency, diet, etc.

1.2 Life cycle of the parasite

Giardia have a simple, biphasic life cycle, interchanging between infective cysts that are excreted into the environment by infected hosts and vegetative trophozoites, which replicate in the intestine. Cysts survive best at low temperature and high moisture conditions (Meyer & Jarroll, 1980) and they have been shown to survive in water for as long as 3 months. Upon ingestion and passage through the stomach, which induces excystation, the trophozoites adhere reversibly to the mucosal epithelium of the small intestine and multiply exponentially by a process of binary fission (Adam, 1991). The process of encystation is poorly understood, its induction *in vitro* by bile (Gillin *et al.* 1987, 1989) remaining controversial (Lujan *et al.* 1998). Variable numbers of trophozoites differentiate into cysts in the lower reaches of the small intestine and in the bowel, and these can be excreted in large numbers. *Giardia* cysts are highly infective, as shown by infectivity trials with human volunteers which indicated that as few as 5-10 cysts were sufficient to transmit infection (Rendtorff, 1954).

1.3 Recognised species in the genus *Giardia*

Giardia belong to the phylum Sarcomastigophora, Class Zoomastigophorea and are classified in the Order Diplomonadida, whose members are all binucleate (Sogin *et al.* 1989;

van Keulen *et al.* 1993; Hashimoto 1994, 1995; Kulda & Nohýnková, 1996). Filice (1952) defined three morphological types within the genus: *Giardia agilis*, found in amphibia; *Giardia muris*, found in rodents, birds and in reptiles; and *Giardia intestinalis* (syn *G. lamblia*, *G. duodenalis*), in mammals (in which it predominates), birds and reptiles. The use of scanning electron microscopy has led to the description of two new *Giardia* species, *G. psittaci* and *G. ardeae* (Erlandsen & Bemrick, 1987; Erlandsen *et al.* 1990). *Giardia microti* is another species, which has been identified by Feely (1988) on the basis of cyst morphological differences, and is found in rodents.

Throughout this thesis, *G. intestinalis* will be used to describe *G. lamblia* and *G. duodenalis*. All *Giardia* from humans and large mammals belong to the *G. intestinalis* group. Within this group, isolates from different host species are largely indistinguishable on the basis of trophozoite size, cyst size and other morphological criteria. *Giardia intestinalis* trophozoites are approximately 10 to 12 μm long and 5 to 7 μm wide and they have two nuclei. The cytoskeleton consists of a median body, ventral disk and four pairs of flagella. A Golgi-like apparatus and rough endoplasmic reticulum has been reported in encysting trophozoites (McCaffery & Gillin, 1994) and perhaps in rudimentary form in vegetative trophozoites (Lanfredi-Rangel *et al.* 1999). Other eukaryotic organelles (e.g. mitochondria, peroxisomes, smooth endoplasmic reticulum, and nucleoli) have not been identified in the parasites.

1.4 Host Specificity and transmission

Survival in fresh water and an ability to establish infections from minute doses has given *Giardia* a reputation as the most common cause of epidemic waterborne diarrhoeal disease in North America. LeChevallier *et al.* (1991) reported that more than 80% of surface water samples from 66 sites in North America contained *Giardia* cysts. The relative importance of humans, animals and birds to surface water contamination is unknown. The

contribution of wildlife to zoonotic infection of humans is an important unresolved issue that has particular relevance to public health and environmental management authorities (Erlandsen, 1994).

In several cross-transmission studies undertaken between 1970 and 1988, the infectivity of *Giardia* isolates from humans and various animal species was tested in other animal hosts (Davis & Hibler, 1979; Goltz, 1980; Faubert *et al.* 1983; Wallis *et al.* 1984; Erlandsen *et al.* 1988). There was a surprisingly large variability in the results from different laboratories. Many of the studies were characterised by what can now in hindsight be seen as design deficiencies (e.g. inadequate controls, use of animals which had not been proven *Giardia*-free, assessing susceptibility to infection for animals that passed cysts for only a day), which leave in doubt some of the conclusions on cross-species transmission. More recent studies have revealed that *Giardia* from humans can differ substantially in genotype (discussed below), and it is now apparent that genetic differences between parasite samples may influence significantly the outcome of infectivity trials involving different host species. Nash *et al.* (1987) and Visvesvara *et al.* (1988) demonstrated the variable infectivity of different human isolates experimentally and concluded that the variability in published experimental data was due to the existence of genetically distinct isolates, or strains, of *Giardia* that are part of a zoonosis. The success of experimental transmission was a major determining factor for the World Health Organisation to recommend that giardiasis be considered a possible zoonosis.

1.5 Host responses to infection and antigenic properties of the parasite

In humans, the incidence of *Giardia* infection is known to be highest throughout childhood, showing a decrease during adolescent years to adult levels (Oyerinde *et al.* 1977). Infections usually induce strong antibody responses, including secretory IgA directed against both surface and internal trophozoite and cyst antigens (Mayrhofer & Waight-Sharma, 1988;

Waight-Sharma & Mayrhofer, 1988a; Chaudhuri *et al.* 1992; Rosale-Borjas *et al.* 1998). Severe, prolonged giardiasis occurs in hypogammaglobulinaemic patients (Vinayak *et al.* 1987), which suggests that a humoral antibody response is important for protective immunity. In studies of the *Giardia*-specific immune response in experimentally infected volunteers, IgM, IgG and IgA antibodies were detected in the serum of 100%, 70% and 60% of the volunteers respectively, whilst IgA antibodies were detected in 50% intestinal samples (Nash *et al.* 1990c). That antibodies contribute to protective immunity is also suggested by the partial resistance of rats that were challenged with *G. intestinalis* trophozoites after receiving bile collected from convalescent donor rats that had been infected with the same isolate (Mayrhofer & Waight-Sharma, 1988). Rats that received bile from uninfected controls were not protected. Other experiments using mutant *xid* mice (deficient in B cells) and immunoglobulin-deficient mice showed that these animals were unable to control *G. muris* infections (Snider *et al.* 1985, 1988). Furthermore, an important role of secretory IgA in eliminating of trophozoites from the intestinal habitat has been indicated in studies using natural and experimental hosts (Farthing, 1990; Stäger & Müller, 1997). The role of secreted variant-specific surface protein (VSP) H7-specific IgA antibodies from the milk of pre-infected mothers was investigated in a murine system (Stäger *et al.* 1998). In offspring infected with a cloned line of *G. intestinalis* (GS/M-83-H7, which produces VSP H7), an immediate appearance of new VSP was observed among the trophozoites in the intestine after ingestion of VSP H7-specific antibodies. *In vitro* studies have also shown that VSP H7-specific IgA antibodies aggregated and were cytotoxic for trophozoites expressing the VSP H7 antigen (Stäger *et al.* 1998).

The importance of T cells in host immunity to *G. intestinalis* infection is evident from experiments in which both neonatal athymic (nude) and severe combined immunodeficient (SCID) mice were found unable to resolve infections (Gottstein & Nash, 1991). Earlier studies with *G. muris* (Heyworth *et al.* 1987; Stevens *et al.* 1978) had shown that nude mice,

as well as mice that had been injected with CD4-specific antibodies, suffered prolonged infections. More recently *G. muris* and *G. intestinalis* infections were examined in mutant mice that were deficient in one or two immune system components (deficient in B cells, αβ- or γδ-T cells, T and B cells, signal transducer and transcription-factor-6, γ-interferon or interleukin-4), using wild-type mice as controls (Singer & Nash, 2000a). In contrast to previous results, they reported that whilst B cell-deficient mice were able to eliminate the majority of parasites within 3 weeks of infection, αβ-T cell deficient mice and animals treated with CD4-specific antibody were unable to control the disease even 4 weeks post infection. They concluded that a T cell-dependent process is required to control acute *Giardia* infections and that it is independent of antibody and B cells. However the weight of evidence appears to indicate that in normal immunocompetent hosts, humoral responses are important in protective immunity to giardiasis and also in controlling the disease during chronic infections (Stäger &Muller, 1997). Defensins, which are small cysteine-rich cationic proteins produced by Paneth cells in the small intestine, may also play a role in non-immune host defence (Aley *et al.* 1994). In a recent study, normal bacterial flora in the intestinal tract have been implicated to play an important role in protecting against *G. intestinalis* infections (Singer & Nash, 2000b).

1.6 Novel characteristics of *Giardia*

Although *Giardia* have the distinction of being the first protozoa to be described (by van Leeuwenhoek in 1681 – Dobell, 1932), it is only in the past 20 years that they have been identified as novel organisms that differ in many characteristics from conventional eukaryotes. On the basis of ultrastructural morphology (Kabnick & Peattie, 1990) and phylogenetic analysis of ribosomal RNA (Sogin *et al.* 1989; van Keulen *et al.* 1993) and polypeptide (Hashimoto *et al.* 1994, 1995) sequences, *Giardia* are now considered to represent one of the earliest known branches of the eukaryotic evolutionary tree. A number of

features show them to resemble bacteria more than eukaryotes, eg. the absence of introns from all characterised *Giardia* genes, bacterial-like ribosomal RNA and short rRNA intergenic spacers (Healy *et al.* 1990; Upcroft *et al.* 1997), and extremely short mRNA 5' untranslated ('leader') sequences although these still possess eukaryotic characteristic (Kirk-Mason *et al.* 1989; Holberton & Marshall, 1995). Many transcripts in *Giardia* appear to be uncapped (Yu *et al.* 1998; Knodler *et al.* 1999) and putative ribosome binding sites within the coding sequence have been described (Yu *et al.* 1998; Garlapati *et al.* 2001). However, mRNA in *Giardia*, like that in higher eukaryotes, is polyadenylated and most of the characterised structural genes possess a conventional eukaryotic polyadenylation signal sequence (AGTRAA) downstream from the stop codon (Peattie *et al.* 1989; Adam *et al.* 1991; Svärd *et al.* 1998; Ey *et al.* 1999).

1.7 DNA and chromosome content of *Giardia*

The cell biology of *Giardia* remains poorly defined. For example, what is the role of the two nuclei? Do they contain identical genetic information? What is the mechanism of translation initiation, what is the number of chromosomes, and the significance of, gene rearrangements and genome instability? What is the mechanism of gene regulation and antigenic variation? In view of the many reported short 5' mRNA leader sequences, do the mRNA's possess an internal ribosome entry site (IRES) as observed for picorna viruses (Chen & Sarnow, 1995; Yu *et al.* 1998)? Many these questions still need to be addressed properly.

The *G. intestinalis* genome has a high GC content with different estimates of 42% (Nash *et al.* 1985), 46.8% (using $C_o t$ analysis, Boothroyd *et al.* 1987) and 48% (using buoyant density, Ortega-Pierres *et al.* 1990). Preliminary analysis of sequences from the *Giardia* genome project (http://www.mbl.edu/Giardia - see Section 1.8) has yielded a more reliable estimate of 46%, similar to these previous estimates (Smith *et al.* 1998).

The ploidy of *Giardia* has been difficult to determine. Kabnick & Peattie (1990) studied the incorporation of [^{3}H]-uridine and showed that both nuclei are transcriptionally active, and by using *in situ* hybridisation that each nucleus contained rRNA genes in similar amount. However, the distribution of the chromosomes between the two identically sized nuclei is unknown. The actual gene content of the two nuclei, i.e. whether they contain identical or complementary sets of genes and exchange genetic material remains unclear, despite recent rigorous studies (Bernander *et al.* 2001).

In other protozoa, e.g. *Plasmodium* and *Toxoplasma*, a pattern of distribution of chromosome size-classes (the karyotype) which is characteristic of each particular species has been identified (Sheppard *et al.* 1989; Sibley & Boothroyd, 1992). In contrast, *Giardia* chromosomes cannot be easily separated and identified by cross-field gel electrophoresis techniques. Co-migration of some chromosomes and frequent chromosomal rearrangements (observed in cloned isolates) has made it difficult to describe a complete or stable karyotype for *G. intestinalis* (Adam *et al.* 1988a; Le Blancq *et al.* 1991b). On the basis of close sequence similarities that were found between large regions of different chromosome-sized DNA molecules, Adam *et al.* (1988a) raised the possibility of diploidy or polyploidy in the nuclei of *Giardia*. Detection by allozyme analysis of apparently heterozygous isolates of *Giardia* has also indicated that the organisms are at least diploid (Andrews *et al.* 1989; Meloni *et al.* 1995).

Different techniques have been used to determine the number of chromosomes. Adam *et al.* (1988a) identified 4-5 major chromosome bands and a number of minor bands by using pulsed-field gel electrophoresis (PFGE). Using field inversion gel electrophoresis (FIGE) and counter-clamped homogenous gel electrophoresis (CHEF) techniques, 7 to 8 different chromosome bands were detected by Upcroft *et al.* (1993). A number of variable "minor" chromosomes were also identified which are presumably duplications of most or all of the major chromosomes and different copy numbers of rDNA repeats were suggested to account

for differences in the size of homologous chromosomes (Adam *et al.* 1988; Adam, 1992). Korman *et al.* (1992) also used PFGE to separate chromosome-sized DNA molecules from 22 isolates of *G. intestinalis*. They identified nine bands per isolate, with different staining intensities indicating that some bands might represent more than one chromosome. However using confocal microscopic examination, five chromosome-like bands were detected in trophozoites undergoing binary fission (Erlandsen & Rasch 1994). Karyotype and Southern analysis of eight *G. intestinalis* isolates using 13 genetic markers identified five chromosomes with stable cores and variable sub-telomeric regions (Le Blancq & Adam, 1998). The number of chromosomal molecules in *Giardia* remains uncertain, with estimates ranging from 8 to 50 per trophozoite (Kabnick & Peattie, 1990: Adam *et al.* 1988). Ribosomal RNA (rRNA) genes have been identified near the telomeric repeats located near the ends of the chromosomes (Adam *et al.* 1991; Le Blancq *et al.* 1991a).

1.8 The *Giardia* genome database (*http://www.mbl.edu/Giardia*)

Because of the impact of *Giardia* on human and animal health and the putative pivotal position of these organisms in early eukaryote evolution, a collaborative project involving several research groups (M. L. Sogin, S. Aley, R. Adam, G. Olsen, F. D. Gillin, H. G. Morrison) was established in the mid 1990's to sequence the genome of an axenic isolate of *G. intestinalis* (isolate WB, clone 6). The strategy involves the construction of plasmid libraries containing size-specific genomic DNA inserts derived from fragments generated either by partial cleavage of the DNA with the restriction endonuclease *Tsp*509 I, or by random shearing. Primers that anneal to the (vector) T3 or T7 polymerase promoters have been used in sequencing reactions to characterise each insert from both flanks. Phagemid and cosmid libraries have also been constructed to obtain sequence information on larger inserts and facilitate compilation of the sequences into larger contiguous sequences ('contigs'). Many potential genes have been identified already (Smith *et al.* 1998; McArthur *et al.* 2000).

The Genome Project is expected to reveal new insights about the early evolutionary history of *Giardia* and other eukaryotes, perhaps enabling better definition of stages that occurred in the evolution of eukaryotic cells from prokaryotes.

Large-scale sequencing of the *Giardia* genome should also provide detailed information about the organisation and character of coding and non-coding regions, and about chromosome structure. It is already yielding a wealth of detail on many structural genes, especially single- or low-copy loci, for which the assembly of randomly-acquired sequence data should be relatively straightforward. However, it remains to be seen whether sequence data from the large vsp gene family can be assembled from random sequence runs with the same degree of certainty. This is an important aspect, addressed in this thesis.

1.9 Propagation of *Giardia in vivo* and *in vitro*

Most *Giardia* isolates do not grow well in adult experimental animals and only a few host species have been found useful for experimental purposes. *Giardia muris* are naturally infective for mice. Although successful growth of a few *G. intestinalis* isolates in adult mice has been reported (Müller & Stäger, 1999; Singer & Nash, 2000), most grow poorly. Gerbils have been a more successful animal host for *G. intestinalis* in some laboratories (Nash *et al.* 1988; Visvesvara *et al.* 1988). However, the importation of these rodents into countries such as Australia is banned. More recently, suckling mice have been used to initially establish isolates and to propagate different subtypes of *G. intestinalis*, including some that have proved refractory to axenic culture (Mayrhofer *et al.* 1992; 1995; Monis *et al.* 1998). The first successful short-term propagation method of *G. intestinalis* trophozoites *in vitro* was reported by Karapetyan in 1960. However, it was not until 1976 that a method for propagating trophozoites in long-term axenic cultures was described (Meyer, 1976). This was an important step, as it enabled isolates to be 'immortalised', made them available to various laboratories as reference isolates for biochemical, metabolic and pharmacological studies,

and for the first time, enabled investigators to establish cloned lines. The modified (bile supplemented) TYI-S-33 medium (Keister, 1983) is now widely used for propagating *G. intestinalis* by axenic *in vitro* culture. However, knowledge about the nutritional requirements of *Giardia* remains rudimentary and a defined medium of known composition for culturing the parasites has not been developed yet. Portland 1 (American Type Culture Collection [ATCC] # 30888), the first isolate established in axenic culture by Meyer in 1976 from a patient in Portland, Oregon), and the WB isolate (ATCC # 30957), derived from a patient who acquired giardiasis in Afghanistan and established in culture by Smith *et al.* (1982), have been used as prototype cultures throughout the world. The bulk of published genetic, biochemical and immunochemical data on *Giardia* is based on these two isolates.

Isolates of *G. intestinalis* differ in their ability to grow *in vitro* (Meloni & Thompson, 1987; Andrews *et al.* 1992; Binz *et al.* 1992), with a much higher rate of success in the establishment of isolates from humans or livestock than from dogs (Meloni *et al.* 1992; Monis *et al.* 1998). Even among axenic cultures, significant growth rate differences have been observed, with evidence that these may reflect innate genetic differences (Meloni & Thompson, 1987; Andrews *et al.* 1992; Binz *et al.* 1992; Karanis & Ey, 1998). The effects on trophozoites of long-term *in vitro* growth are unknown. In the absence of natural selective pressures e.g. host immunity, and the necessity to encyst, it is conceivable that long-term cultures may acquire (or lose) by genetic drift phenotypic characteristics that would reduce their survival and fitness *in vivo*. Infections involving multiple subtypes of *G. intestinalis* have been reported by several groups (Andrews *et al.* 1989; Upcroft & Upcroft, 1994; Carnaby *et al.* 1995) and from the data available, it seems that mixed infections are not uncommon. Isolation techniques clearly exert important selective influences on the organisms that are made available for genetic analysis. Isolates that comprise a mixture of genotypes can yield cultures in which one or another subtype predominates, depending on

whether the organisms are propagated in suckling mice or by axenic culture (Andrews *et al.* 1992; Binz *et al.* 1992; Karanis & Ey, 1998).

1.10 Comparative studies on isolates of *Giardia*

It is now apparent, from the results of genetic analyses during the past 15 years, that the apparent morphological homogeneity of *G. intestinalis* isolates masks a substantial genetic diversity. Immunological and molecular genetic studies have resolved discrete genetic subtypes within *G. intestinalis*.

The use of antisera, raised against specific isolates to examine a large sample of axenic isolates and clones, allowed isolates initially to be divided into three broad groups (1, 2, 3) based on their profile of reactivity with the different antibodies (Nash & Keister, 1985). Further studies, using a variety of different techniques, supported the authenticity of these three groupings. Comparison of chromosomal DNA banding patterns of 14 different isolates of *G. intestinalis* by Southern blot analysis (using recombinant plasmids containing *Giardia* DNA as probes) revealed one common banding pattern in six isolates and unique banding patterns for some isolates (Nash *et al.* 1985). The results of PCR amplification and sequence analysis of a rRNA gene segment from a variety of *G. intestinalis* (Weiss *et al.* 1992), confirmed the validity of the three previously defined subgroups in *G. intestinalis* (Nash, 1992; Nash & Mowatt, 1992a).

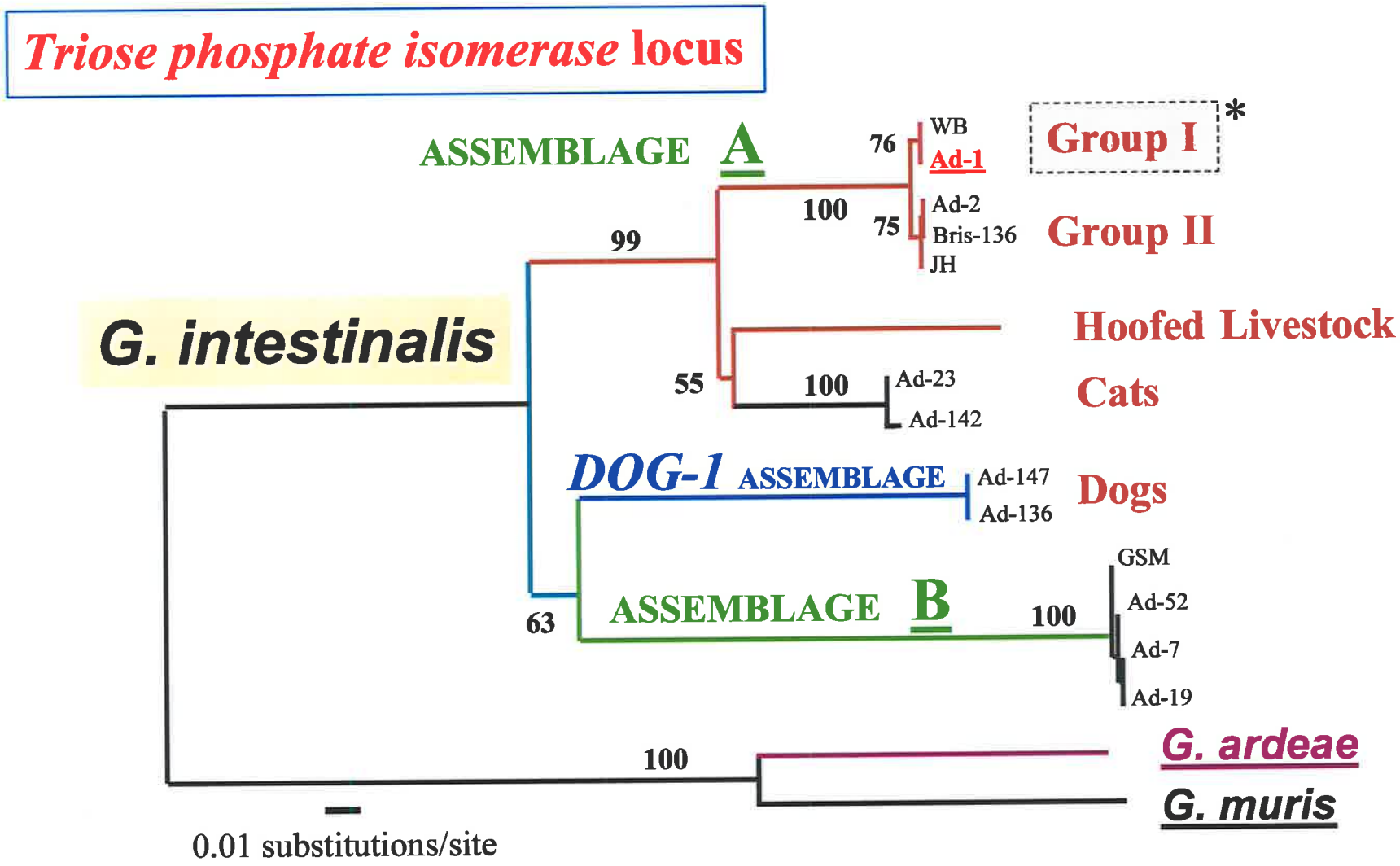
Analysis of polymorphisms detected by isoenzyme electrophoresis provided initial indirect evidence of interspecies polymorphisms. Significant genetic diversity between isolates of *G. intestinalis* was shown by Meloni *et al.* (1988) who examined 15 enzymes and defined 13 banding patterns ('zymodemes') from 30 isolates. More rigorous evidence to support the existence of genetic subgroupings was obtained by allozymic analysis of isoenzyme electrophoresis data for 26 enzymes in *G. intestinalis* isolates from Australia (Andrews *et al.* 1989). Allozymic analysis of multilocus enzyme electrophoresis data

subsequently defined two major genetic assemblages (A and B), which include all tested *Giardia* isolates from humans (Mayrhofer *et al.* 1995). Concurrent molecular genetic studies which compared isolates for restriction fragment length polymorphisms (RFLP) within specific loci by hybridisation analysis of genomic DNA or RFLP and partial nucleotide (nt) sequence analysis of DNA amplified in PCR from genomic DNA, confirmed and extended these findings (Homan *et al.* 1992; Nash & Mowatt, 1992a; Ey *et al.* 1992, 1993a,b, 1996; Monis *et al.* 1996; Baruch *et al.* 1996; Lu *et al.* 1998). Using data acquired from vsp genes (*tsa*417, *tsp*11, *vsp*1267) as well as more conserved loci encoding the enzymes glutamate dehydrogenase, triosephosphate isomerase and CTP synthetase, the translation initiation factor, Elongation Factor 1-α, and 18S ribosomal RNA, it has been shown that Assemblage A and B (corresponding respectively to groups (1 + 2) and group 3 of Nash) represent two deeply rooted genetic lineages within *G. intestinalis* (Baruch *et al.* 1996; Monis *et al.* 1996; 1998, 1999; Swarbrick *et al.* 1997). These studies also revealed (1) that assemblage A and B genotypes include many isolates from animals, some involving zoonotic infections; (2) the existence of distinct subtypes (defining allozymic groups I and II, which correspond to groups 1 and 2 of Nash) within Assemblage A and significant genetic heterogeneity within Assemblage B (Nash group 3); and (3) the existence of apparently host-specific subtypes in dogs, hooded livestock and possibly cats and other animals (Nash, 1992; Ey *et al.* 1992, 1993a,b, 1996, 1997; Monis *et al.* 1996, 1998, 1999). On the basis of these findings, *G. intestinalis* must be considered to be a species-complex rather than a single species (**Fig. 1.1**). The axenic Ad-1/c3 *G. intestinalis* isolate used in this project belongs to genetic group I of Assemblage A (subtype A-I).

1.11 Antigenic variation and variant-specific surface proteins (VSP)

Giardia are extremely successful parasites, infecting all known species of vertebrates and causing occasional epidemic infections in human populations (Juranek, 1979). The

Figure 1.1. Defined subtypes of *Giardia intestinalis* (*G. lamblia*). Adapted from Fig. 81, P.T. Monis, Ph.D. Thesis, Adelaide University (1997). The relationships between representative isolates of each lineage were inferred by Neighbour Joining analysis of nucleotide sequences, using Tamura-Nei distances (Kumar et al. 1993) determined for a 480-bp segment of the triosephosphate isomerase locus (Monis et al. 1999). Branch lengths are proportional to mean differences and bootstrap support (%, calculated from 500 iterations) for the major branches is indicated. *Giardia ardeae* and *G. muris* were used as outgroups to define the *G. intestinalis* complex. Designated genetic groups within *G. intestinalis*, e.g groups I and II of Assemblage A, are indicated. The Ad-1 isolate, which was used for most of the work described in this thesis, belongs to genetic group I (*), abbreviated 'Subtype A-I'.


Triose phosphate isomerase locus
ASSEMBLAGE A
76
WB
Ad-1
Group I
*
100
Ad-2
75
Bris-136
JH
Group II
99
G. intestinalis
Hoofed Livestock
55
100
Ad-23
Ad-142
Cats
DOG-1 ASSEMBLAGE
Ad-147
Ad-136
Dogs
GSM
63
ASSEMBLAGE B
100
Ad-52
Ad-7
Ad-19
100
G. ardeae
G. muris
0.01 substitutions/site

factors that influence infectivity and the development of clinical disease are multifactorial (Farthing, 1992). However, one notable characteristic of the parasites that may prolong infections is antigenic variation, a phenomenon observed for a number of other medically important parasitic protozoa such as African trypanosomes. During the course of an infection, trypanosomes sequentially produce different variant surface glycoproteins (VSGs) that are encoded by an extensive genomic repertoire of genes estimated at more than 1,000 (Thon *et al.* 1989). However, there is only a single active VSG transcription unit, which ensures the expression of only one VSG at any one time. The early VSGs (expressed early in the early stages of infections) are encoded by genes located near the telomeres (Myler *et al.* 1984; Laurent *et al.* 1984). Non-telomeric genes are expressed later. These genes are flanked by sequences homologous to those present in expression sites, but the extent of these homologies varies. In the later stages of infections, new hybrid VSG genes may arise as a result of recombination between pseudo- and functional genes (Roth *et al.* 1986; Thon *et al.* 1989). This process of antigenic variation causes variants to appear within infected individuals at a frequency that prevents elimination of the entire parasite population as immunity develops against the dominant parental phenotype, resulting in chronic infections (Vanhamme & Pays, 1995).

In *Giardia* infections, early observations of strong agglutination antibody responses and surface immunofluorescence indicated that trophozoites possessed dominant surface antigens (Nash *et al.* 1985, 1988). These molecules were shown by cell surface radio-iodination and immunoprecipitation experiments to be major polypeptide components of the plasma membrane (Smith *et al.* 1982; Nash *et al.* 1983; Einfeld & Stibbs, 1984; Nash & Keister, 1985; Edson *et al.* 1986; Kumkum *et al.* 1988; Nash *et al.* 1988). Most of these studies were carried out between 1983 and 1991 by Nash and colleagues, using a panel of monoclonal antibodies raised against trophozoites from subclones of different axenised isolates of *G. intestinalis* (Adam, 1991; Nash, 1992; Nash & Mowatt, 1992b). Label-fracture

electron microscopy and immunocytochemistry studies using monoclonal antibodies raised against some of these predominant surface antigens revealed that the trophozoites were completely covered by a surface 'coat', comprising any one of a group of related proteins - termed *variant-specific surface proteins*, or VSP (Pimenta *et al.* 1991). These proteins were found to constitute a unique family of cysteine-rich proteins that varied in molecular mass from 33 to 200 kDa and contained multiple copies of the divalent cation-binding 'Cys-Xaa-Xaa-Cys' motif (Adam *et al.* 1988b, Adam, 1991, 1992; Gillin *et al.* 1990), where Xaa may be any amino acid. Adam *et al.* (1988b) were the first to report the cloning and characterisation of a segment of a vsp gene, *crp*170 – subsequently renamed *vsp*A6 (Yang & Adam, 1994). This gene was identified from a *Giardia* genomic DNA library by screening bacterial colonies for the presence of antigen using a monoclonal antibody (6E7) that was specific for an epitope in the tandem repeat region of this protein (Mowatt *et al.* 1994). This mAb was reported in a subsequent study to be cytotoxic for trophozoites that produced VSP A6 (Adam *et al.* 1992). The progeny of the trophozoites that survived exposure to mAb E6 were surface-labelled with ^{125}I and found, by immunoprecipitation and SDS-PAGE analysis, to express different surface antigens.

In a parallel study, Aggarwal and Nash (1988) showed that antigenic variation in *Giardia* is not an artefact of *in vitro* culture. They infected gerbils with *G. intestinalis* trophozoites from clonal cultures whose VSP were known to react with particular mAbs. Within 7 days of infection, the trophozoites recovered from the intestines of these animals no longer reacted with the clone-specific mAb, indicating that they no longer expressed that particular antigen. Similar experiments conducted on human volunteers revealed a gradual loss of the parental trophozoite surface antigen within two weeks of infection, with replacement by a variety of different surface proteins (Nash *et al.* 1990). Studies on axenic cultures have shown that trophozoites continuously release their surface antigens (VSP) into the medium (Nash *et al.* 1983; Papanastasiou *et al.* 1996a).

1.12 General structure of Variant-Specific Surface Proteins

Comparisons of the polypeptides inferred from the nucleotide sequences of characterised vsp genes from different isolates of *G. intestinalis* have confirmed features that were considered characteristic of these proteins on the basis of findings from the first 2-3 genes that were examined (Adam *et al.* 1988b, 1992; Gillin *et al.* 1990; Mowatt *et al.* 1991; Adam, 1991). These properties include:

1. An N- terminal hydrophobic segment which resembles the signal peptide ('leader') sequences of known secreted proteins from other eukaryotes and bacteria. This presumably directs the translocation of the nascent polypeptide into the cisternal space of the endoplasmic reticulum. The "mature" form of VSP417-1 (TSA417, synthesised by the WB/C6 clone of *G. intestinalis*) has been purified and found by N-terminal sequence analysis to lack the signal peptide segment (Aley & Gillin, 1993). This is probably removed by signal peptidase as occurs for secreted proteins in other organisms.
2. A variable segment, comprising the bulk of the polypeptide. This can differ in length and amino acid composition. On the basis of the inferred amino acid sequence of this segment, two types of VSP can be distinguished: those for which this segment is comprised largely of multiple tandem repeats and those that lack tandem repeat elements (**Fig. 1.2, Table 1.1**).
3. A highly conserved, 29-residue hydrophobic segment near the C-terminus. This is predicted to form a membrane-spanning α-helix, anchoring the protein in the plasma membrane.
4. An invariant, 5-residue C-terminal segment (-CRGKA$_{COOH}$) which may extend from the inner face of the plasma membrane into the cytoplasm.
5. These features are illustrated in **Fig. 1.2.**

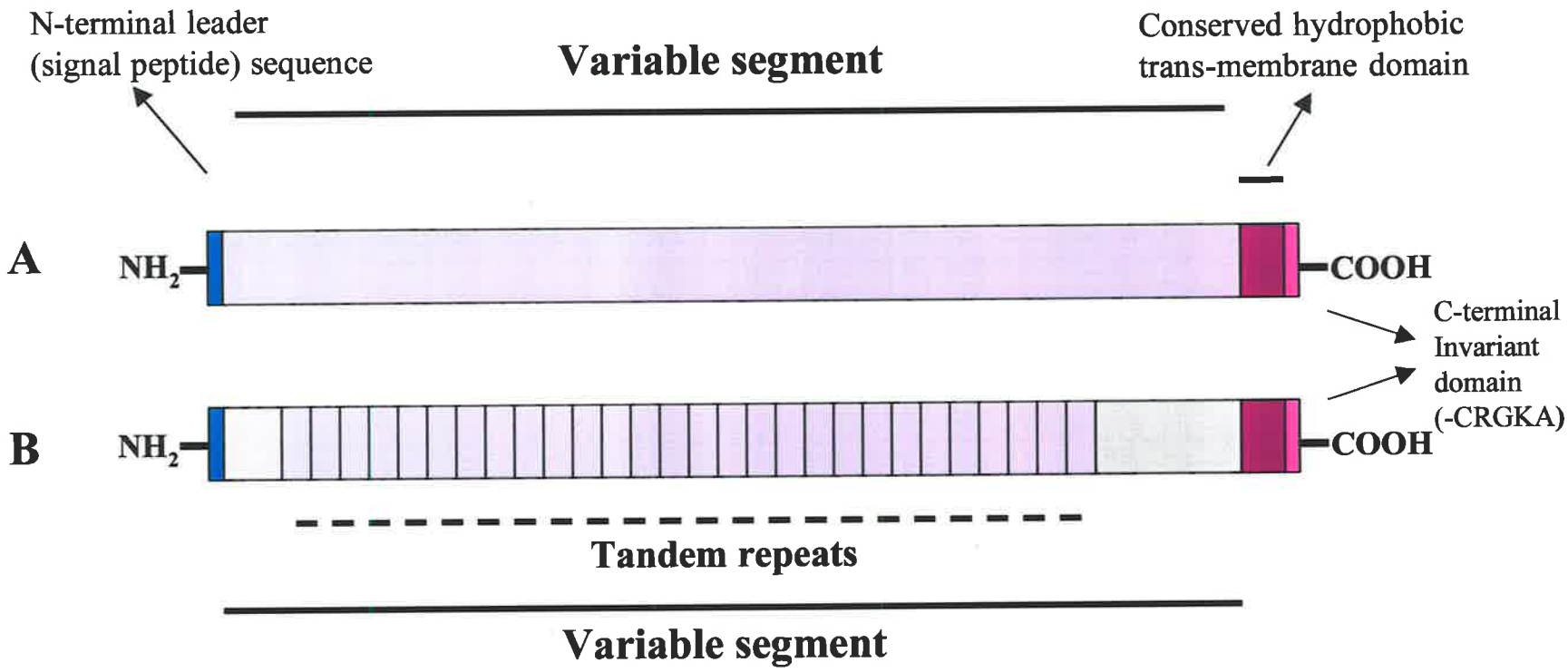


Figure 1.2. Schematic representation of a *Giardia* variant-specific surface protein (VSP). The signal peptide, variable segment and conserved C-terminal domains are indicated. Two types of VSP are depicted, those lacking (A) or containing (B) tandem repeat elements.

Table 1.1. Known *G. intestinalis* VSP genes

VSP gene	Identified in		Tandem repeat		References
	Isolate(s)	Genotype(s)	Length,	Copy no.	
*vsp*A6.1 (*crp*170)	WB-A6	AI	195 bp,	18-23×	Adam et al., 1988b
*vsp*A6.2 (allele)	WB-A6, WB-1269	AI	195 bp,	9×	Yang & Adam 1994
*vsp*A6.3 (allele)	WB-A6, WB-1269	AI	195 bp,	8×	Yang & Adam 1994
*vsp*A6-S1	WB-A6	AI	195 bp,	>1×	Yang & Adam 1995
*vsp*A6-S2	WB-A6	AI	201 bp,	4×	Yang & Adam 1995
*crp*136	WB1-B	AI	120 bp,	23.5×	Chen et al., 1995
*crp*65	WB1-B-M3	AI	228 bp,	4.2×	Chen et al., 1996
*vsp*52	Ad-1	AI	111 bp,	12.5×	Ey et al., unpubl.
*vsp*C5	WB-C5	AI	105 bp,	26×	Yang et al., 1994
*vsp*C5-S1,S2	WB-C5	AI	Nd		Yang & Adam 1995
*vsp*417-1 (*tsa*417)	WB, Ad-1 & others	AI, AII	-		Gillin et al., 1990, Ey et al., 1993c, 1996; Ey & Darby 1998
*vsp*417-2 (*tsp*11)	Ad-1 & others	AI, AII	-		Ey et al., 1993, 1996, unpubl.
*vsp*417-3	Bris-136 & others	AII	-		Ey et al., 1998
*vsp*417-4	Ad-1, Bris-136 & others	AI, AII	-		Ey et al., 1999
*vsp*417-5 (=*vsp*417-1)	Bris-136 & others	AII	-		Ey & Darby, 1998
*vsp*417-6	Ad-1, Bris-136 & others	AI, AII	-		Ey et al., unpubl.
*vsp*417-7	Ad-1, Bris-136 & others	AI, AII	-		Ey et al., unpubl.
*vsp*1267	WB-1267	AI	-		Mowatt et al.,1991
*vsp*1269 (*crp*72)	WB-A6, WB-1269	AI	-		Adam et al., 1992
*vsp*4A1	O2-4A1	AI/II-like	-		Papanastasiou et al., 1997
*vsp*H7	GS/M-H7	B	-		Nash & Mowatt., 1992b
*vsp*H7-1	GS/M-H7	B	-		Nash et al., 1995

Nd, not determined.

1.13 The VSP repertoire

Studies on the rate of antigenic variation and the size of the VSP repertoire have been reported by Nash and colleagues (1990, 1992b), using two isolates of *G. intestinalis* (WB and GS/M) representing group 1 (subtype A-I) and group 3 (subtype B) respectively. Three mAb were used to detect antigenic variants within clones of the WB line and a fourth mAb (G10/4) was used to detect cells bearing the corresponding target epitope within GS/M clones. The rate of appearance (in 'non-expressing' clones) of cells (variants) bearing the epitopes recognised by these mAbs was measured, after establishing for each isolate that the growth rates of the parental phenotype and the emerging antigenic variants were similar. The single epitope studied in the GS/M isolate was re-expressed more frequently than those detected in the WB clones, leading the authors to conclude that the isolates differed in their switching rate. They also concluded that the two isolates possessed different sized VSP repertoires. Their estimates for the switching frequency (from the expression of one surface antigen to another) were estimated at one in every 6 cell divisions (for GS/M) or 1 in every 12-13 divisions (WB isolate), from which they concluded that the parasites had an antigen repertoire of 64-150. However the conclusions depend on whether mAb G10/4 (used with the GS/M-H7 clone), as well as the other mAbs, cross-react with other VSP. If their target epitopes were not unique, this would result in a more frequent appearance of (say) G10/4-positive variants and lead to an erroneous estimate of the switch rate for 'this' particular VSP. The method used by Nash and colleagues to calculate their estimates was not clearly explained. Although no other studies on VSP switch rates have been published, it seems likely that the frequency of appearance of a particular variant would be determined by both the size of the vsp gene repertoire and the switching rate. Future studies aimed at improving the accuracy of these estimates should couple the detection of variants identified by a larger range of mAb or, preferably by *in situ* mRNA hybridisation using probes that have been proven to be locus-specific, with the development of mathematical models.

In a subsequent study, Nash & Mowatt (1992b) used an oligonucleotide probe that was specific for the conserved 3' end of vsp genes (that part encoding the hydrophobic trans-membrane segment) to undertake a quantitative spot-blot hybridisation analysis of genomic DNA from the WB isolate. The radiolabelled probe was tested for hybridisation to spots containing serial dilutions of known quantities of WB genomic DNA or a plasmid construct containing the *vsp*H7 gene. The amount of radioactivity was quantified following high stringency washes. Assuming a genome complexity of 3×10^7, it was estimated that the WB genome had a vsp gene repertoire of 133-151 (Nash & Mowatt, 1992b). This estimate may include both functional and nonfunctional (pseudo) vsp genes, in contrast to the detection of expressed VSP by mAb in the aforementioned studies.

1.14 Vsp genes and evidence of replicated loci

Antigenic variation, defined as changes in exposed antigenic determinants at frequencies much higher than can be accounted for by natural rates of mutation, appears to be one of the major means by which some parasitic protozoa maintain persistent infections in the face of host-protective immune attack. As the highly immunogenic VSP are the major antigens on the surface of *Giardia* trophozoites, they are potentially very important in the development of host-protective immunity. The capability of trophozoites to change the antigenic character of their surface coat, i.e. to 'switch' to other VSP, seems likely to be advantageous for the parasites by prolonging infections through the successive replacement of parental populations by their variant progeny.

In *Plasmodium falciparum*, genes that encode a family of surface proteins (rifin) have been found to comprise up to 7% of the protein-encoding genes in the genome (Gardner *et al.* 1999). In comparison, for *G. intestinalis*, initial estimates have indicated that VSP-encoding and related (pseudo gene) sequences may occupy as much as 2.4% of the genome (Smith *et al.* 1998)). The function(s) of *Giardia* VSP may still be uncertain, but the occupation by vsp

gene-like sequences of such a significant portion of the genome indicates that the encoded proteins are vital for the parasites' survival.

It has been known for some time that more than one copy of some vsp genes exist in the genome. For example, two identical copies of the *vsp*1267 gene (which encodes the VSP detected on trophozoites in clone 1267 of the WB isolate) were identified in genomic DNA extracted from this clone (Mowatt *et al.* 1991). The two genes were arranged in a tail-to-tail configuration and separated by a 3-kb intergenic region. Copies (tentative alleles) of another vsp gene (*vsp*A6, syn. *crp*170), which was first identified as an mRNA transcript (Adam *et al.* 1988b) encoding a VSP containing 23 copies of a 65-amino acid repeat, have also been identified and partially characterised (Yang & Adam 1992). These apparently non-expressed alleles (*vsp*A6.2 and *vsp*A6.3 respectively) possess only 8 or 9 tandem copies, respectively, of the same repeat unit and differ from *vsp*A6 by only 8 nt substitutions within the characterised segments of their coding regions.

Trophozoites that survived treatment of the WB/A6 clone with the cytotoxic VSP A6-specific mAb produced different VSP and surprisingly, the *vsp*A6 allele was not detected in DNA extracted from these resistant lines (Adam *et al.* 1992; Yang & Adam, 1994). The apparent loss of the expressed allele (*vsp*A6) had important implications, as it suggested that vsp gene loss might be a frequent occurrence, perhaps indicative of recombinations leading to the creation of new genes. The reported loss of this gene helped to initiate studying on vsp gene stability in this laboratory (Ey *et al.* 1993a-c, 1996, 1997, 1998, 1999).

In 1994, the *vsp*C5 gene (containing 26 tandem copies of a 105-bp repeat) was characterised (Yang *et al.* 1994). Two related genes, *vsp*C5-S1 and *vsp*C5-S2 identified in a subsequent study (Yang & Adam, 1995), but only partially sequenced, were very similar to *vsp*C5 within their 5'-coding and upstream flanking regions. However, they both diverged abruptly from *vsp*C5 in the first (5' proximal) repeat unit. These findings indicated that recombination and duplication events can contribute to vsp gene evolution.

There is no reliable estimate on the number of non-expressed (non-functional, or pseudo) vsp genes that may exist in the *G. intestinalis* genome, nor whether genetic recombinations result in occasional or frequent re-activation of such loci to functionality. This is a complex but important issue, as it has relevance to questions about vsp gene stability and repertoire size. Only superficial attention has been given to this subject in the published literature on *Giardia*. At present, little is known about how the vsp gene family evolved and how it is maintained, or about the stability of individual genes or vsp gene subfamilies.

1.15 Vsp gene expression / regulation

Telomeric expression–linked sites are believed to be responsible for antigen (VSG) switching in African trypanosomes (Pays *et al.* 1994). However with the exception of one report (Upcroft *et al.* 1997) of a metronidazole-resistant mutant subclone of the WB isolate which contained a vsp gene transcript that was different from that of the (non-resistant) parent line - possibly due to the development (selection) of metronidazole resistance in an antigenic variant, there is so far no evidence for a similar mechanism of expression in *Giardia*.

In an earlier report, Chen *et al.* (1995) characterised a transcript of *crp136*, a vsp gene that they found was expressed at high levels by trophozoites in a subclone (WB-1B-M3) of the WB isolate that had been sublethally irradiated with ultraviolet light and then selected for resistance to metronidazole. The gene encodes a 136-kDa protein containing 23 copies of a 40-amino-acid repeat. The authors proposed that the tandem peptide array encoded by the *crp*136 repeats possessed sequence homology with a class of toxins (sarafotoxins) produced by a burrowing adder, emphasising that the initial symptoms of giardiasis (stomach cramps, nausea, vomiting and diarrhoea) are similar to those of adder-bite victims. This analogy seems trivial, since the effects of snakebite are largely the result of acute systemic

(haemolytic) anaphylaxis, whereas those of giardiasis are chronic symptoms attributable to interference (by the parasites) of digestive processes and mucosal epithelial function. The level of amino acid sequence similarity between the CRP136 repeat and the pharmacologically active sarafotoxin (which is derived by proteolytic cleavage of a larger tandemly arrayed repeat within a precursor polypeptide) is marginal and its significance seems dubious.

Using the heterogenous transcript (of isolate WB-1B-M3) as a probe to screen a genomic library derived from the same isolate, Chen *et al.* (1996) identified another gene, *crp65*, that encoded a deduced 65-kDa cysteine-rich protein. This gene contained a long tandem repeat segment and exhibited significant homology with *crp*136 (95.8% and 84.8% nt sequence identity in the non-repeat 5' and 3' coding regions, respectively). However, the repeat unit of *crp*65 was distinct from that of *crp*136. On the basis of an apparent sequence similarity between the *crp65* repeats and epidermal growth factors (EGF), which participate in protein-protein interactions, Chen *et al.* (1996) proposed that CRP65 may interact with certain types of host proteins. They proposed that *crp*136 and *crp*65 belonged to a vsp gene subfamily, the members of which are characterised by the possession of tandem repeats within a highly conserved 'cassette' (Chen *et al.* 1996). Using 'chromosome walking' to sequence longer tracts of genomic DNA, *crp*136 and *crp*65 were identified near the telomeric regions of chromosomal DNA from the WB1-B-M3 clone. Two other non-vsp genes situated upstream from each of these vsp genes were characterised and identified as encoding a protein kinase and an ankyrin homologue. These segments were proposed by the authors to be mobile gene clusters that move within the genome and among the telomeres on different chromosomes (Upcroft *et al.* 1997). They also considered that the expression of *crp*136 and *crp*65 by the resistant line was causally linked to, ie. underpinned, metronidazole resistance (Upcroft *et al.* 1997). However, as mentioned earlier there is no evidence to indicate that the expression of either gene was not coincidental. With ample evidence of antigenic variants, it

seems more likely that metronidazole resistance arose via enzymic mutants within the progeny of variant trophozoites that had already 'switched' to express *crp*136 or *crp*65.

1.16 Tandem repeat sequences in VSP

The polypeptides inferred from available vsp gene nucleotide sequence data (**Table 1.1**) can be divided into two distinct groups, on the basis of whether or not they possess tandem repeat regions. Several of the 12-16 characterised vsp genes possess large segments that are comprised of tandem repeats. The occurrence of one of these genes, *vsp*A6 (*crp*170), in different clones and isolates has been examined by testing for VSPA6-like proteins (using mAb) and for *vsp*A6-like nucleotide sequences by Southern hybridisation analysis of genomic DNA (Nash & Mowatt, 1992b; Mowatt *et al.* 1994). As mentioned earlier, non-expressed alleles of *vsp*A6 have been detected in the WB/A6 genome. Yang & Adam (1994) characterised vsp genes possessing different number of tandem repeats, but the existence or stability of these loci/alleles in other isolates of similar or different genotype remains unclear. It seems likely, however, on the basis that these genes share highly conserved or similar segments, they would be susceptible to recombination.

Of all the vsp genes that have been characterised to date, those comprising the *vsp*417 subfamily have been studied in most detail – expressly to address the questions about their stability and evolution. Isolates representing different subtypes of *G. intestinalis* have been examined for the presence or absence of specific loci belonging to this subfamily and where homologous loci have been identified, how these have diverged over an evolutionary time scale (Ey *et al.* 1993a-c, 1996, 1997, 1999; Ey & Darby, 1998; Ey & Darby, unpublished data). The *vsp*417 subfamily is based on the prototype vsp gene, *tsa*417 (redesignated *vsp*417-1), first described in the type A-I (Assemblage A/group I isolate, WB, by Gillin *et al.* 1990). Five genes, all closely related to *vsp*417-1, have been identified as stable loci in genomic DNA from all additional type A-I (Nash group 1) isolates that have been tested.

Paralogs of most of these loci have also been identified in type A-II (Nash group 2) isolates, which represent a closely related but distinct sublineage of *G. intestinalis*, as well as the in more divergent 'Hoofed livestock' genotype (**Table 1.1, Fig. 1.1**). The results of these studies indicate that most of the loci that comprise the *vsp417* gene subfamily are very stable, both as intact entities (full-length coding sequences) and in their overall structure. They have survived a long period of evolution, with evidence of various highly conserved segments that reflect selective pressures on the encoded VSP (**Fig. 1.3**). These loci show evidence that they have evolved as a result of ancestral gene duplications and subsequent mutational divergence. The contribution of homologous recombination to the evolution of these genes has still to be evaluated.

1.17 The biological function(s) of VSP

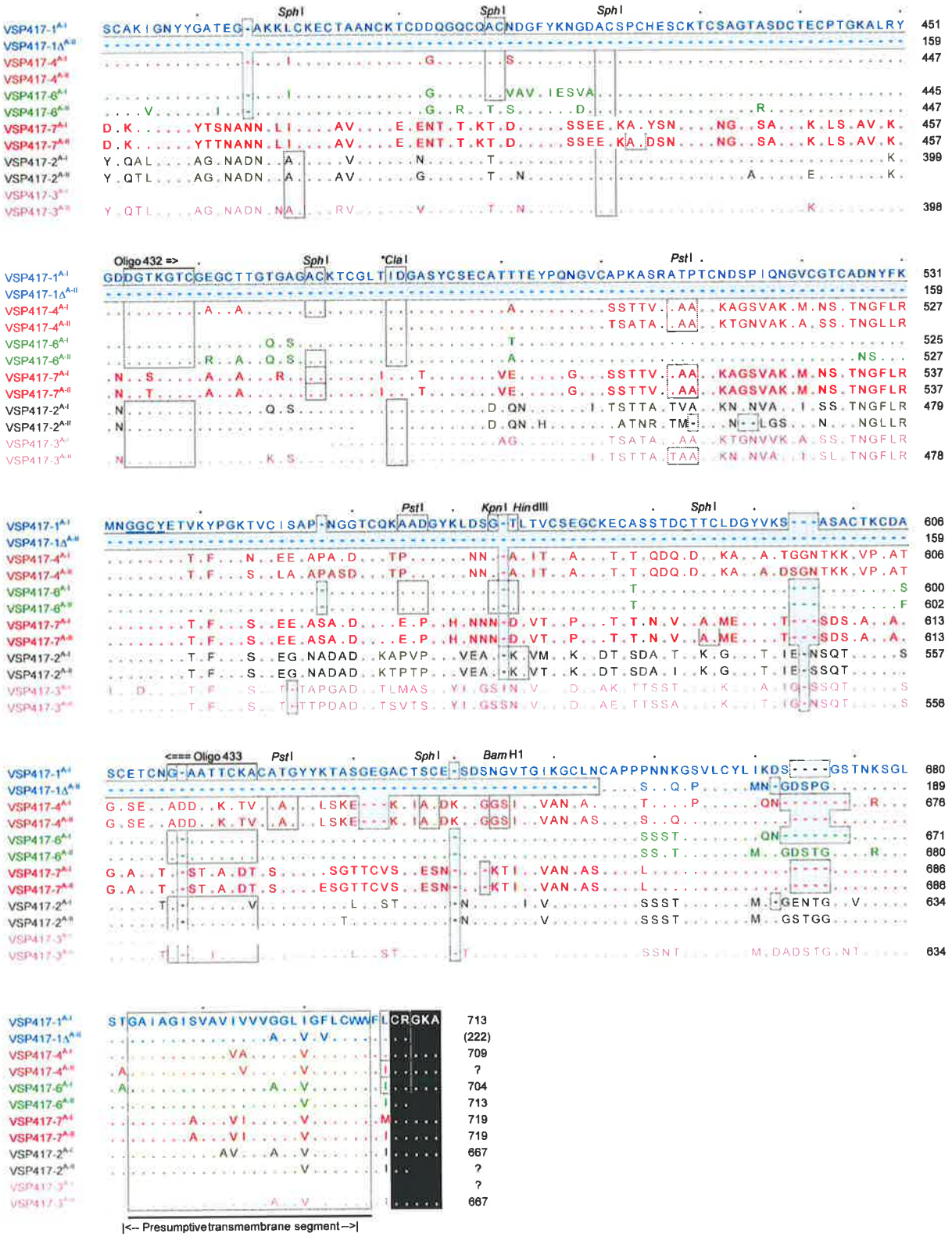
Whether VSP have biological functions other than to form a protective coat over the surface of the trophozoites, thereby shielding them from the lytic or degradative action of intestinal bile salts and proteases is unknown. That vsp gene sequences comprise a large portion of the relatively small *G. intestinalis* genome indicates their importance for the parasite. However, it is possible that this reflects only that these proteins are necessary to maintain cell viability within the intestine and that host immunity, targeted predominantly against the coat protein, has resulted in the evolution of a parasite that has the capacity to switch its coat protein on a regular basis. The observations of strong agglutinating antibody responses against the surface antigens support such an hypothesis. Additional functions of the VSP, e.g. sequestration of extracellular metal ions (Zn^{2+}, Fe^{2+}) via the CXXC motifs or as stage-specific differentiation (encystation) 'antigens', cannot be dismissed at present.

The susceptibility of *G. intestinalis* trophozoites to intestinal proteases (and the importance of the VSP coat in conferring resistance to these enzymes) was investigated by Nash *et al.* (1991). They exposed trophozoites from three cloned lines (WB-2X, GS/M-H7,

Figure 1.3. Aligned amino acid sequences inferred from characterised genes belonging to the *vsp417* subfamily in type A-I and type A-II *Giardia intestinalis* (overleaf). The comparison includes homologues (VSP417-1$^{A\text{-}I}$ and VSP417-1Δ$^{A\text{-}II}$, VSP417-2$^{A\text{-}I}$ and VSP417-2$^{A\text{-}II}$, VSP417-3$^{A\text{-}I}$ and VSP417-3$^{A\text{-}II}$, VSP417-4$^{A\text{-}I}$ and VSP417-4$^{A\text{-}II}$, VSP417-6$^{A\text{-}I}$ and VSP417-6$^{A\text{-}II}$, VSP417-7$^{A\text{-}I}$ and VSP417-7$^{A\text{-}II}$) that occur in isolates of A-I or A-II genotype respectively. The 682-residue VSP417-2$^{A\text{-}I}$ sequence is corrected from the originally reported 667-residue polypeptide (Ey et al. 1993c) by replacement of residues 77--79 ('MSS') in the original TSP11 sequence with the 18-residue segment, 'CIGANFFMYKGGCYDKEK'. Residue numbers, commencing with the N-terminal methionine, are indicated at right. Gaps [--] introduced to maximise the alignments are shaded and boxed. Dots indicate amino acid identity between VSP417-1$^{A\text{-}I}$ (TSA 417, Gillin *et al.* 1990; row 1) and corresponding positions in the other proteins. Positions corresponding to diagnostic restriction sites within the parent genes are boxed and labelled (above) with the relevant endonuclease, as are those corresponding to PCR primers 2144, 432, 433 and 731. For each protein, the predicted N-terminal signal peptide and probable site of cleavage is shown, together with the presumptive membrane-spanning hydrophobic segment (boxed, dotted borders) and invariant hydrophilic segment ('-CRGKA', highlighted) at the C-terminus. The alignment comprises only part-published data of Dr. P.L. Ey, who kindly provided the figure.

Fig. 1.3. Aligned amino acid sequences inferred from characterised genes belonging to the *vsp417* subfamily in type A-I and type A-II *Giardia intestinalis*. Unpublished data of Dr. P.L. Ey.

Fig. 1.3 (continued)



B6) of *G. intestinalis* to α-chymotrypsin and trypsin, each at concentrations up to 20 mg/ml (Nash et, al., 1991). Each clone expressed a unique VSP. After 24 hours incubation with the proteases, a significant decrease in the number of surviving WB-2X and B6 trophozoites was observed whereas GS/M-H7 trophozoites (an Assemblage B genotype, **Fig. 1.1**) showed no significant susceptibility (cytotoxicity) to either enzyme (Nash *et al.*, 1991). *Giardia* clones that produce different VSP may therefore differ in their susceptibility to intestinal proteases. Other studies on TSA417 (from WB/C6 trophozoites) and VSP 4A1 (from isolate O2-4A1) have indicated that these VSP are relatively resistant to proteolysis by trypsin and chymotrypsin, both as purified polypeptides and within the surface coat on viable trophozoites (Gillin *et al.* 1996; Papanastasiou *et al.* 1997). However, under reducing conditions (in the presence of dithiothreitol), trypsin and chymotrypsin resulted in complete breakdown of VSP 4A1 (Papanastasiou *et al.* 1997). Intrachain disulfide bonding between cysteine residues is likely to be important in VSP folding (Aley & Gillin, 1993), stabilising the proteins and conferring resistance (to the VSP and thus the cells) against digestive enzymes and detergents in the gut.

The potential of VSP to function as Zn^{2+}-binding proteins is uncertain. Intestinal enzymes and also nucleic acid-binding proteins and metallo-enzymes require Zn^{2+i} ions to be functional (Coleman, 1992). Most of the inferred VSP sequences contain a segment related to 'zinc finger' motifs (Nash & Mowatt, 1993) which resembles the LIM, and RING-finger motifs found in many zinc-binding proteins from higher eukaryotes (Sanchez-Garcia & Rabbitts, 1994; Freemont *et al.* 1991). Both LIM- and RING-finger motifs occur in DNA-binding proteins, many of which are believed to act as transcriptional factors. These, together with the multiple copies of CXXC motifs, seem likely to account for the reported ability of the polypeptides to bind Zn^{2+} and other metal ions (Zhang *et al.* 1993; Nash & Mowatt, 1993). Competition between the VSP on *Giardia* trophozoites and Zn^{2}-dependent intestinal (host) enzymes for trace amounts of Zn^{2+} may resulted in reduced enzyme activities and

nutritional malabsorption by the host. This would depend, however, on dissociation of bound Zn^{2+} ions from the enzymes. The ability of three purified VSP (VSP 417-1, VSP H7 and VSP 4A1) to bind Z^{2+} ions have been examined in independent studies (Zhang *et al.* 1993; Nash & Mowatt, 1993; Papanastasiou *et al.* 1997). In each case the proteins were denatured and then tested for the capacity to bind radioactive Zn^{2+} ions. VSP 417-1 (TSA 417) and VSP H7 were found to bind Zn^{2+} ions after renaturation in the presence of ß-mercaptoethanol. However the binding was not specific to Zn^{2+}, as competition by other metal ions, e.g. Fe^{2+} and Ca^{2+}, was observed. Papanastasiou et al (1997) examined the ability of denatured and reduced VSP 4A1 to bind Zn^{2+} ions. Consistent with the other experiments, upon removal of these reducing agents, the 'blotted' VSP 4A1 was able to bind Zn^{2+} ions. However, only a 1:1 molar ratio of Zn^{2+} to protein was detected in immunoprecipitated, detergent-solubilised 'native' VSP that had been purified from trophozoites grown in the presence of Zn^{2+}, indicating that the native proteins may not bind significant amounts of Zn^{2+} ions (Papanastasiou *et al.* 1997).

In a study designed to determine whether *Giardia* trophozoites have an essential need for Zn^{2+} ions, Nash & Rice (1998) tested 34 zinc-finger-active compounds for toxicity against two axenic *G. intestinalis* clones, WB /1269 (tested by *in vitro* culture) and GS/H7 (tested in infected adult mice). Twenty-nine of the 34 drugs exhibited activity against *G. intestinalis in vitro* and one of the most active compounds, disulfiram, was also active in infected mice. These experiments demonstrated the vital importance of Zn^{2+} ions for the parasites' growth and survival. However, they provide no evidence for the involvement of VSP in binding these ions, as the demonstrated toxicity may be the result of inactivation or inhibition of enzymes and/or transcription factors within the cells.

1.18 Background to the project

The focus of this laboratory has been to identify and characterise the various loci that comprise the *vsp*417 gene subfamily in type A-I and type A-II *G. intestinalis*, to elucidate

how rapidly (and in what manner) these genes are evolving, and to investigate the size and nature of other vsp gene subfamilies. One clonal line, Ad-1/c3, had been examined in detail. A polyclonal rabbit antiserum (R102) and a murine mAb (2D4) had been raised against the dominant VSP. The R102 antiserum was raised against the purified surface antigen - solubilised from trophozoite membranes in Triton X-100 and fractionated, in the presence of the detergent, by size-exclusion and ion-exchange chromatography. The 2D4 mAb was raised by immunising mice with whole Ad-1/c3 trophozoites (Mayrhofer & Ey, unpublished data; Ey *et al.* 1993). These antibodies reacted strongly with >98% of viable cells in the clonal Ad-1/c3 population. Analysis of surface-biotinylated Ad-1/c3 membranes by SDS-PAGE and Western blotting (using streptavidin-peroxidase) showed three prominent bands, corresponding to polypeptides of approximately 85, 70 and 20 kDa (designated the 'complex').

Work done prior to commencement of this Ph.D. project had indicated that these three polypeptides were associated together non-covalently, as they could not be separated by size, ion-exchange or affinity chromatography. The data therefore indicated either that these proteins were co-synthesised within single trophozoites, or that the smaller (70 and 20 kDa) bands represented breakdown products of the larger polypeptide, similar to the partial proteolytic degradation of TSA417 (VSP 417-1) (Aley & Gillin, 1993). The R102 serum reacted on Western blots with all three polypeptides, whereas mAb 2D4 reacted with viable trophozoites and specifically, on blots, with the non-reduced 85-90 kDa subunit of the complex.

The R102 antiserum had additionally been used to screen an Ad-1 genomic DNA library for antigen expression (Ey *et al.* 1993c). Two reactive colonies (11 and 52) were identified and characterisation of the respective plasmid inserts revealed two vsp genes encoding:

1. A 70-kDa polypeptide (Trophozoite Surface Protein 11, TSP 11) consistent with one of the bands identified by Western blot analysis of Ad-1/c3 trophozoites (Ey *et al.* 1993a). The *tsp*11 gene, since re-designated *vsp*417-2, encodes a polypeptide that has significant amino acid sequence similarity with TSA417 (VSP 417-1) described by Gillin *et al.* (1990). Additional studies have resulted in characterisation of the *vsp*417 gene subfamily in type A-I and type A-II isolates (Ey *et al.* 1996, 1998, 1999; Ey & Darby 1998, unpubl. data).
2. An 89-kDa polypeptide (designated VSP 52) containing 12.5 tandem copies of a 37-amino acid repeat (**Fig. 1.4**) (Ey, Darby, Khanna & Mayrhofer, unpubl. data). This polypeptide exhibits extensive sequence similarity with two other characterised VSP, CRP136 and CRP65 (Chen *et al.*1995, 1996; mentioned earlier) which contain tandem repeats that are unique for each protein and distinct from the VSP 52 repeat. Because the R102 antiserum (raised against the purified Ad-1/c3 surface antigen 'complex') had identified separate bacterial colonies that harboured plasmid constructs containing distinct genomic fragments encompassing the *vsp*417-2 (*tsp*11) and *vsp*52 genes, this had supported the previous immunological data which indicated that VSP 417-2 and VSP 52 (with deduced molecular masses of 69 kDa and 85 kDa respectively) were the larger subunits of the 'complex' present on the surface of Ad-1/c3 trophozoites.

At the commencement of this project, mRNA extracted from Ad-1/c3 trophozoites was reverse-transcribed into complementary DNA (cDNA) and analysed using the polymerase chain reaction (PCR), with the aim of detecting the presence of transcripts of *vsp*417-2 and *vsp*52 within the population by the amplification of specific, known segments of each gene using locus-specific PCR primers. This yielded products that were characteristic of *vsp*417-2, but none that corresponded to *vsp*52 using forward primers that were specific for sequences within the 5' region of the latter gene. However, DNA corresponding in size to an expected segment in the 3' portion of *vsp*52 was amplified in good yield using primers

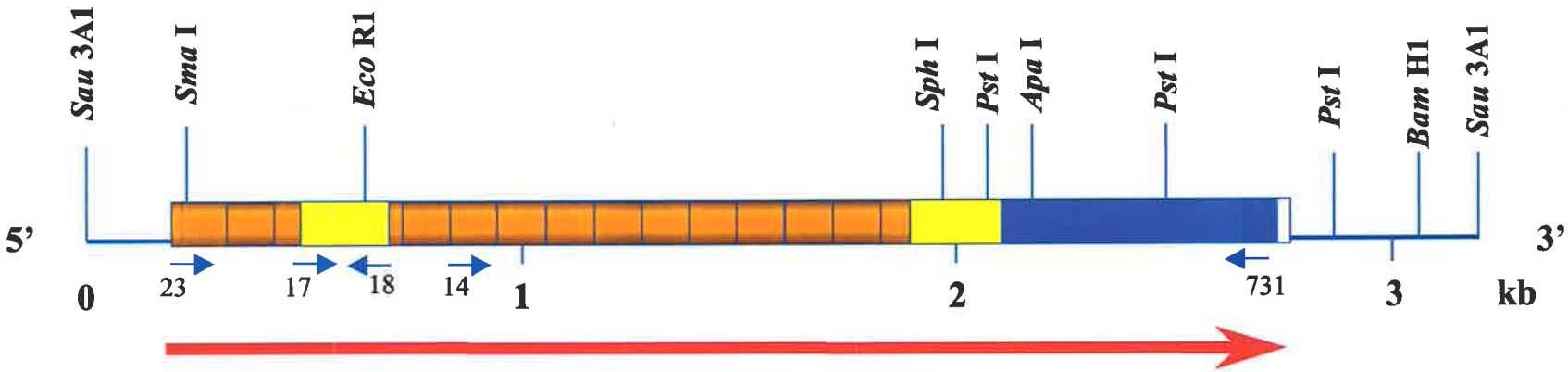


Figure 1.4. Restriction map of the *vsp52* gene. Tandem repeats are shown by red shaded boxes. Two homology-block regions are indicated as yellow boxes. The 3' segment of the gene is shown as a dark blue box. Key restriction sites are indicated. Short arrows represent annealing sites for PCR primers (numbered) which were used to amplify segments by PCR or for sequencing.

that annealed to sites within this segment of the gene. This product was cloned and analysed by nt sequence determination, which revealed that it was a segment of a distinct but nevertheless related vsp gene, one that contained a tandem repeat region also. Subsequently, *in situ* mRNA hybridisation analysis of Ad-1/c3 trophozoites (of which 98% were strongly labelled after incubation with fluorescein-labelled 2D4 mAb or with R102 antibodies) confirmed the RT-PCR results, i.e. that *vsp*52 was not expressed by the 2D4 mAb-binding majority of cells within the cultures, but in the contrary by only a very small minority (<1%) of the cells. Surprisingly, cells expressing the *vsp417*-2 (*tsp11*) were also a very minor subpopulation (<1%) within these cultures, indicating that neither VSP 417-2 nor VSP 52 were components of the Ad-1/c3 VSP complex produced by the majority of the trophozoite population. These results showed that the R102 antiserum was not specific for the immunising polypeptides, but that it cross-reacted with an unknown number of other VSP.

1.19 Aims of the project

The initial aim of my project was to examine Ad-1/c3 trophozoites for particular vsp gene transcripts related to the *vsp*417-2 and *vsp*52 genes that had been identified using the R102 anti-(Ad-1/c3 VSP complex) antiserum, with the intention of learning more about the antigenic and structural properties of these proteins. At the time this study was commenced, data obtained in this laboratory had already indicated the existence of a stable subfamily of vsp genes (encoding TSA417-like polypeptides [discussed earlier; Fig. 1.3]) and provided insights on conserved structural motifs (possibly important for polypeptide folding) and how the genes encoding these proteins have evolved. However, little comparative information was available about other vsp genes or about the existence of other vsp gene subfamilies. The *Giardia* genome project, which even now provides only fragmentary sequence data from randomly cloned inserts, had not commenced at this time. We were interested in exploring other gene subfamilies, especially in view of the discovery, at the start of the project, that

neither of the two vsp genes (*vsp417*-2 and *vsp52*) identified using the R102 anti-(Ad-1/c3 VSP complex) serum were components of the complex. The broad aims, therefore, were to:

1. Examine other vsp gene subfamilies within the genome of Ad-1/c3 *G. intestinalis* clone.
2. Investigate the structural relationships between the VSP encoded by loci belonging to related subfamilies.

In general terms, it was anticipated that the accumulation of data on additional vsp genes in an isolate that was independent of the WB isolate (on which the *Giardia* Genome project is based), but of the same subtype (A-I), would improve our knowledge about vsp gene stability and expression. We considered that the analysis and comparison of physically separated, characterised cloned genomic inserts, especially those containing very similar (replicated) vsp loci, might have unique value - particularly if the number and similarity of the sequences prevents unambiguous compilation of the random, overlapping sequences obtained from the Genome project. Specific goals were, therefore, to:

a) Identify and characterise new vsp gene loci within the Ad-1/c3 genome.

b) Compare these with known loci that have been described in the literature, or with any that may be identified by similarity searches of sequences from the *Giardia* Genome database.

c) Examine the loci for possible regulatory elements (conserved upstream sequences) and recombination sites (breakpoints in the coding and noncoding flanking regions).

d) Investigate, by *in situ* mRNA hybridisation using specific anti-sense probes, the expression of particular vsp genes within trophozoite populations.

Chapter 2

Materials & Methods

2.1 Bacteria and plasmid cloning vectors

The *E. coli* strain DH5αF' was used in this study. Three cloning vectors, pBluescript SK+ (Stratagene, La Jolla, CA.), pGEM-7Zf(+) and pGEM-3Zf(+)(Promega Corp), were used to clone DNA fragments of interest. They all confer ampicillin resistance, allowing selection of transformed bacteria.

2.2 Chemicals

Proteinase K, restriction endonucleases, digoxigenin-11-dUTP, alkaline-phosphatase-labelled anti-digoxigenin antibodies, calf intestinal alkaline phosphatase, *Tth* DNA polymerase, T4 DNA ligase and other molecular biology reagents were purchased from Boehringer-Mannheim, New England Biolabs, Promega Corporation, or Pharmacia Biotechnology Ltd. RNase A and herring sperm DNA were obtained from Sigma Chemical Corporation. Solutions of RNase A (10 mg/ml) were heated for 30 min at 100°C to inactivate residual DNase activity. Other reagents were either of molecular biology or analytical grade.

2.3 Buffers and solutions

Alkaline Phosphatase Staining solution: This was used for blots or *in situ* hybridisation experiments. Digoxigenin labelled probes were detected by incubating blots or slides in the dark with blotting buffer containing 4-nitroblue tetrazolium (7.5 mg/ml), phenazine methosulphate (5 mg/ml), and 5-bromo-4-chloro-3-indolyl-phosphate (5 mg/ml). This staining solution was prepared fresh and used immediately. Ten-fold concentrated stocks of 4-nitroblue tetrazolium (75 mg/ml in 70% dimethylformamide), phenazine methosulphate (50 mg/ml in 70% dimethylformamide) and 5-bromo-4-chloro-3-indolyl-phosphate (50 mg/ml in dimethylformamide) were stored at –20°C.

Blotting buffer: 0.1 M Tris-HCl, 25 mM diethanolamine-HCl, 0.1 M NaCl, 2 mM $MgCl_2$, 1 µM $ZnCl_2$ (pH 9.55)

Ampicillin: Added to broth and solid media at a final concentration of 50 µg/ml.

Electrophoresis marker dye (10× Blue Juice): 15% (w/v) Ficoll 400, 0.3% (w/v) bromophenol blue, 0.1 mg/ml RNase A. Stored at 4°C. Diluted 1:10 for use as a tracking dye in electrophoresis of DNA.

Buffer 1: 0.1 M Tris-HCl, 0.15 M NaCl (pH 7.5).

Denhardt's solution (100×): 10 g of polyvinyl pyrrolidine, 10 g of bovine serum albumin, and 10 g of Ficoll 400 per 500 ml.

DIG-dNTP mix (for incorporation of digoxigenin-11-dUTP during PCR):

dATP, dCTP, dGTP (2 mM each), 1.3 mM dTTP and 0.7 mM digoxigenin-11-dUTP.

Lysis buffer: 100 mM Tris-HCl, 50 mM NaCl, 50 mM EDTA, 1% SDS (pH 8.0).

Nutrient agar: Blood base agar (Difco), prepared without the addition of blood.

Nutrient broth (Difco): Prepared at double strength (16 g/l), with added NaCl (5 g/l).

PE (Phosphate-EDTA) buffer, 10×: 1.33 M sodium phosphate, 10 mM EDTA (pH 6.9). One litre contained 93.75 g of $NaH_2PO_4.2H_2O$, 261.5 g of $Na_2HPO_4.2H_2O$ and 3.72 g of $Na_2EDTA.3H_2O$.

Physiological saline: Sterile 0.15 M NaCl.

PCI (Phenol/chloroform/*iso*-amylalcohol):

Phenol:chloroform:*iso*-amylalcohol, 50:48:2 by volume.

PBS (Phosphate buffered saline): 100 ml of sterile 0.132 M sodium phosphate (pH 7.4), plus 400 ml of sterile physiological saline.

PBSm (Phosphate buffered saline (modified, 20×): KCl (4 g), anhydrous Na_2PO_4 (23 g), anhydrous KH_2PO_4 (4 g) and NaCl (250 g) per litre, autoclaved. The osmolarity of the

working solution was 420 mOsm/kg, suitable for maintaining the viability of *Giardia* trophozoites.

Pre-hybridisation buffer: 20× SSC (6 ml), deionised formamide (12 ml), 0.2 M sodium-phosphate, pH 6.4 (2 ml), 50× Denhardt's solution (2.5 ml) and 10 mg/ml herring sperm DNA (0.2 ml). The final solution has a composition equivalent to 5.3× SSC, 5.5× Denhardt's and 53% formamide.

SSC (Sodium-saline-citrate) buffer, 20×): 3 M NaCl, 0.3 M trisodium citrate.$2H_2O$ (pH 7.0).

TBE (Tris-Borate-EDTA) buffer, 20×): 0.9 M Tris base, 0.9 M boric acid, 20 mM disodium EDTA (pH 8.3).

Tris-EDTA (TE) buffer: 10 mM Tris-HCl, 0.1 mM EDTA (pH7.4).

Water: Unless otherwise specified, this was always deionized and autoclaved.

2.4 Synthetic oligodeoxynucleotides

The oligodeoxynucleotides used are listed in **Table 2.1** and were purchased from GeneWorks (Adelaide, Australia). Potential PCR primers were designed and assessed for Tm, %GC and false properties, (ability/inability to form hairpins and primer dimers, etc.) using the program Primer (version 2.0, Scientific and Educational software). This program provided a useful guide, but it had only limited utility as it is unable to analyse primers containing degenerate sites. Specific primers were designed as required, using sequence data acquired during the course of this study. Degenerate primers were constructed by comparing codon-based nucleotide sequence alignments in conjunction with the corresponding (inferred) amino acid sequence alignments for the relevant loci.

Table 2.1. Oligodeoxynucleotide primers used in this study

Primer	F/R[a]	Sequence[b]	Identity to
1	F	5'-CGGGATCCGA GCTGAGAAGG ATTTTAATG-3'	*vsp*417-2
2	F	5'-CGGGATCCATGGCAAAGTGTACTACGCAGA-3'	*vsp*417-2
3	F	5'-GGAATTCTATCAGTTTTGTCCATCCAAGAT GA-3'	*vsp*417-2
4	F	5'-AGCGGATCCATGACTGAGAACTGCAAGGCGTG-3'	*vsp*417-2
6	F	5'-GGGATTCTATCAMCKRCANABGAACCACCARCA-3'	*vsp*417
7	F	5'-CGGGATCCATGGCCTGCACCCAAGAAGCTGAC-3'	*vsp*417-1
8	F	5'-CGGGATCCATGACCCCCGGATGCAAGGCGTG-3'	*vsp*417-1
9	F	5'-CACAAGCTTCTATCACGTAGTTTTGTCGGCAC-3'	*vsp*417-1
14	F	5'- CAACACGCCCAACTGCAA-3'	*vsp*52
17	F	5'-ACCAATGCGTACCTRACTGC-3'	*vsp*52
18	R	5'-GGCTGCATGC CTTACACTT-3'	*vsp*52
25	R	5'- AGTACACGCCGCGCACTTAT-3'	*vsp*417-6
26	F	5'- TGCACGCCTGACACTGAGTT-3'	*vsp*417-6
54	F	5'-CGCCAGGGTTTTCCCAGTCACGAC-3'	M13 primer
55	F	5'-TCACACAGGA AACAGCTATG AC-3'	M13 primer
57	F	5'-TAATACGACT CACTATAGGG CGA-3'	T7 primer
68	F	5'-AACARGAGCG GNCTYWSCRC DGG-3'	VSP 'pre-transmembrane'
80	F	5'-GCGTTTAAA$(T)_{30}$ VN -3'	Oligo (dT)
94	F	5-ACAACGAGAAGGACACATGC-3'	*vsp*52
109	R	5'-TCCGCTATGGTTGCTACGCATG-3'	*Sph* Isite (3'overhang)
120	F	5'-ACGTGCGGCTCAGGCTACA-3'	*vspR2*
121	R	5'- GCTCACGCTGCACTGGTTC-3'	*vspR2*
128	R	5'-TCTTGCACTGCTGGTGGTT-3'	*vspR2*
129	F	5'-ACCCCAARACMGACAACG-3'	*vspR2*
141	F	5'-ACAGTAGTGGTAGTGGTACG	*vsp72*
148	F	5'-GTTGGTATGACTGTCGCTGTG-3'	*vsp72*
149	R	5'-CTTCGTATCAGTCGTACACGTC-3'	*vsp*72p
150	R	5'-TGCATGTCTTTTTGCCACTATC-3'	*vsp*R2
154	F	5'-CCGGAGTTGGTTCAGTAACA-3'	*vsp72*
155	R	5'-CCACATATGCCTTAGCGGT-3'	*vsp72*
156	R	5'-CACAGAGCATGTCACACACC-3'	*vsp72*
167	R	5'-CGTTACAGGCACTGCACACC-3'	*vsp72*
168	F	5'-CTTCCTGAAAGGCAATGGTC-3'	*vp72*
178	R	5'-CGGCTTGCATGTCTCACAGT-3'	*vsp72*

Primer	F/R[a]	Sequence[b]	Homologous to
179	R	5'-GTATCCGTCGAGGCAAGCTG-3'	*vsp72*
183	R	5'-CTGTAGCAGCAGAGCACGTC-3'	*vsp72*
184	F	5'-TGGCTCTGTCACATGCACCA-3'	*vsp72*
190	F	5'-GTTTGTGCAACAGTGCTGGC-3'	*vsp*136
191	R	5'-ACATCACAGCCCTCGACCTC-3'	*vsp*136
192	R	5'-AGGCCGTACACTTATCAGCAC-3'	*vsp*136
193	F	5'-TCCCGGTCGRCGGYTTCT-3'	*vsp72*
194	F	5'-GGGTACTTCCTCTTCATGGG-3'	*vsp72*
195	F	5'-AGTGCRTHYTBTGYTCYGAT-3'	*vsp72*
196	F	5'-GAGTGCATHYTYTGYCATGAT-3'	*vsp72*
197	R	5'-ACYCCCACRTATGCCTTAGC-3'	*vsp72*
198	R	5'-TRCAYAYGAACCACCAGCA-3'	*vsp72*
199	F	5'-TGTTCCAACTGATACCCCTG-3'	*vsp72*
200	R	5'-CCCCATGAAGAGGAAGTACC-3'	*vsp72*
201	R	5'-ATCRGARCAVARDAYGCACTC-3'	*vsp72*
202	R	5'-ATCATGRCARARDATGCACTC-3'	*vsp72*
203	R	5'-TCCCAGTCTTACAAGAGGTGC-3'	*vsp72*
204	R	5'-CCCYTCTGGACACGATGTGC-3'	*vsp72*
205	R	5'-CATGCCGCATGACGTCCAAT-3'	*vsp*136
206	F	5'-GCACAGGCCATCTAGCAGCA-3'	*vsp*136
207	R	5'-GAGGACCGATTCGCATGGAG-3'	*vsp*136
208	F	5'-GACATGTGCCAATGGCTTAG-3'	*vsp*136
731	R	5'-ACGCCTCTAGACYRCANABGAACCACCARCA-3'	VSP 'pre-transmembrane'
969	F	5'-CATCCGCTATGGTTGCTAAG GTAC-3'	*Kpn* I site(3' overhang)
970	F	5'-TTAGCACTCGGATAGGAACTGCA-3'	*Pst* I site (3' overhang)

[a] Forward (F) or reverse (R) primers respectively.
[b] Degenerate primers were made in order to amplify related sequences. Single letter codes stand for: R = A or G; Y = C or T; M = A or C; K = G or T; S = C or G; W = A or T; H = A, C or T; B = C, G or T; V = A, C, or G; D = A, G or T and N = A, C, G or T.

2.5 Axenic culture of *Giardia intestinalis*

Giardia intestinalis trophozoites were grown axenically at 37°C in modified TYI-S-33 medium (Keister, 1983) from cryopreserved stocks in liquid nitrogen.. The composition of

the medium was: Difco bacto yeast extract (10%, autoclaved), 20 ml; ferric ammonium citrate (BDH, 1 g/100 ml), 0.56 ml; sodium bicarbonate (1.4%), 1 ml; sterile newborn calf serum, 20 ml; minimal essential medium powder (Eagle, GIBCO), 130 mg; penicillin, 60 mg; streptomycin sulphate, 100 mg; D-glucose, 2 g; L-cysteine.HCl monohydrate (Sigma), 400 mg; NaCl, 400 mg; ox-bile (Oxide, desiccated), 20 mg; ascorbic acid, 40 mg; anhydrous KH_2P0_4, 120 mg ; anhydrous K_2HPO_4, 200 mg. This solution was made up to 200 ml with water and sterilised by filtration through a 0.22 μm Millipore filter. Cultures were maintained in 13 ml Falcon polystyrene tissue culture tubes and grown to confluency. They were propagated every 2-4 days by removing any sediment of unattached or dead cells whilst the tubes were warm, detaching the trophozoites by cooling each tube for 20-30 min on ice and then using 1 to 2-ml aliquots to inoculate new tubes, which were filled with fresh medium. Trophozoites were collected by centrifuging cold cell suspensions at 220×g for 10 min at 4°C (1300 rpm, Beckman bench-top centrifuge). The cells were resuspended, washed twice in cold PBSm and were counted using a haemocytometer. Unless stated otherwise, the Ad-1 isolate (clone 3, or Ad-1/c3) of *G. intestinalis* was used for all experiments utilising *Giardia* trophozoites or *Giardia* DNA described in this thesis.

DNA extraction procedures

2.6.1 Plasmid DNA isolation: Plasmid minipreps

Plasmid DNA was extracted using either of the following procedures.

Method 1: Plasmid purification was performed by a three-step alkali lysis method (Ish-Horowicz & Burke, 1981). Samples from colonies of interest were mixed with 10 ml of nutrient broth containing 50 mg/ml ampicillin and grown overnight at 37°C. Following centrifugation at 300×g (10 min, 4°C), the cells were resuspended in 1 ml of Solution 1 (50 mM glucose, 25 mM Tris-HCl, 10 mM EDTA, pH 8.0). After 5 min, 2 ml of Solution 2 (0.2 M NaOH, 1% (w/v) SDS) was added and the mixture was incubated for 5-min on ice before

adding 1.5 ml of Solution 3 (5 M potassium acetate, pH 4.8). After a further 5 min incubation on ice, protein, chromosomal DNA, high molecular weight RNA and SDS were removed by centrifugation. The supernatant, containing the plasmid DNA, was mixed with 2 vol of absolute ethanol, left 10 min at room temperature and then centrifuged in a Microfuge (maximum speed, 5 min). The pellet was resuspended in 0.5 ml of 1× Tris-EDTA (TE), transferred to a new tube and incubated with 20 µg of RNase A (37°C, 30 min), followed by an extraction with phenol/chloroform/*iso*-amyl alcohol. DNA in the aqueous phase was precipitated (section 2.7.7) and redissolved in 50 µl of water. Purity and concentration were determined (section 2.7.1).

Method 2: Plasmid DNA was purified using the QIAprep Spin Plasmid kit (QIAGEN, Cat. No: 27104) according to the manufacturer's instructions.

2.6.2 Preparation of genomic DNA from *Giardia* trophozoites

Trophozoites of *G. intestinalis* were grown at 37°C as described (section 2.5). Adherent cells were harvested and washed in ice-cold PBSm. The cells (5×10^7-1×10^8) were resuspended in 0.9 ml of cold 0.1 M Tris-HCl, 0.2 M NaCl, 0.1 M EDTA (pH 8.0). To this was added 100 µl of 10% SDS and 20 µl of 20 mg/ml proteinase K and the mixture (containing 1% SDS) was incubated at 56°C for 3 hrs. The digest was extracted twice with TE-saturated phenol/chloroform/*iso*-amylalcohol and the interface was extracted with 0.5 M Tris-HCl, 0.2 M NaCl; 0.1 M EDTA (pH 8.0) between extractions. The final aqueous layer (1.5 ml) was extracted with chloroform, incubated for 3 hrs at 37°C with 100 µg of RNase A and then re-extracted twice with phenol/chloroform/*iso*-amylalcohol and once with chloroform. The purified DNA was dialysed overnight at 4°C against 2 L of TE and then precipitated at -20°C overnight (section 2.7.7), air dried and redissolved in 100 µl of TE buffer. Purity and concentration were determined from absorbance measurements at 260 nm and 280 nm and by electrophoresis (section 2.7.1).

2.7 Analysis and manipulation of DNA

2.7.1 Agarose gel analysis and quantitation of DNA

DNA samples were subjected to electrophoresis on 1% agarose in Tris-Borate-EDTA buffer. Prior to electrophoresis, 0.1 vol of marker dye (blue juice) was mixed with each sample (Sambrook *et al.*, 1989). Electrophoresis of DNA samples was in horizontal gels containing 1× TBE buffer at 80 V for 2-3 hrs. The gels were stained for 15 minutes in ethidium bromide (2 µg/ml) in distilled water, and then destained in water for 5 minutes. DNA fragments were visualised under UV light and photographed using Polaroid 667 film or with a Tractel Gel Documentation Video System. Total DNA concentrations were determined by measuring the absorbance of solutions at 260 nm (A_{260}) using an ultraviolet DU®-64 spectrophotometer (Beckman Instruments, Fullerton, CA). Calculations utilised an A_{260} of 1 for a 50 µg/ml solution of DNA.

2.7.2 Estimation of DNA fragment lengths

DNA fragment lengths were calculated from electrophoretic mobilities by linear regression analysis using *Eco*RI cleavage fragments of *Bacillus subtilis* bacteriophage SPP-1 DNA as size standards (8.51, 7.35, 6.11, 4.84, 3.59, 2.81, 1.95, 1.86, 1.51, 1.39, 1.16, 0.98, 0.72, 0.48, and 0.36 kilobases (kb) respectively, as listed in the Bresatec Ltd. 1998 catalogue).

2.7.3 Restriction enzyme analyses

DNA samples were incubated overnight at 37°C with 2 units of endonuclease in 20 µl of the appropriate 1× digestion buffer. Reaction mixtures were analysed by electrophoresis as described.

2.7.4 Dephosphorylation of DNA using alkaline phosphatase

To a 30 µl sample of DNA (2 µg), calf intestinal alkaline phosphatase (2 units) and 10× phosphatase buffer (3 µl) were added. The reaction mixtures were incubated for 2 hrs at 37°C, followed by 30 min incubation at 65°C to inactivate the phosphatase. The 5'-dephosphorylated DNA fragment was then gel-purified.

2.7.5 Phosphorylation of DNA

To enable the ligation of DNA fragments to 5'-dephosphorylated vectors, the 5′ ends of the 'insert' fragments were phosphorylated using T4 polynucleotide kinase (5 units) (Pharmacia, Sweden, Cat No. 27-0736-01) in a 10 µl volume containing 66 mM Tris-HCl (pH 7.6), 6.6 mM $MgCl_2$, 10 mM dithiothreitol (DTT), 0.1 mM spermidine and 1 mM ATP. The kinase was then inactivated by heat (65°C, 15 min).

2.7.6 Isolation of DNA fragments from agarose gels

Fragments of DNA that had been fractionated by electrophoresis were extracted from gel slices using the QIAGEN Gel Extraction kit (Cat. No. 28706) according to manufacturer's instructions.

2.7.7 DNA precipitation

DNA was precipitated from solution by the addition of 0.1 vol of 3 M sodium acetate (pH 5.2) and 2.5 vols of absolute ethanol. After mixing and leaving the samples at -20°C for at least 30 min, they were centrifuged in a microfuge at 16,000×g for 20 min. The DNA pellets were washed in cold 70% ethanol, dried *in vacuo* and redissolved.

2.7.8 Extension of restricted DNA from 3′ overhanging ends

This method, described by Upcroft and Healey (1993), was used to add a defined sequence (complementary to an oligonucleotide primer) onto the 3' overhanging ends of suitably restricted DNA fragments, eg. those produced by cleavage with *Pst* I, *Kpn* I or *Sph* I.

The method utilises primers designed to hybridise t a low annealing temperature to 4-5 nt of the 3′ overhanging ends of DNA cleaved by these enzymes (see **Fig. 2.1 A, B**), thereby allowing extension of the 3' end of the DNA using the annealed (generic) primer as the template (see **Fig. 2.1 C&D**). The extended DNA was then suitable for use in conventional PCR, using a combination of a gene-specific primer and the particular generic primer that was used for the extension reaction (**Fig. 2.1 E**). In the PCR step, the generic primer hybridises to the extended (3') complementary ends whilst specificity is imparted by the locus-specific primer. The method enables segments of DNA that are situated between the specific primer site and a nearby restriction endonuclease site to be amplified for analysis (**Fig. 2.1 F**). Controls needed for the PCR using these DNA templates included reactions containing only one of the primer pair, and restricted DNA which had not been 3' extended. The method is outlined in **Fig. 2.1**.

2.7.9 3'-Extension reactions

Samples of *Giardia* genomic DNA (3 µg) were incubated at 37°C overnight with *Kpn* I, *Pst* I or *Sph* I (2 units) in 20 µl reaction volumes. The enzymes were inactivated by heating (65°C, 30 min) or gel purification. Oligonucleotides 969, 970 and 109 (**Table 2.1**) were used as templates for extending the 3' overhanging ends of genomic DNA cleaved by *Kpn* I, *Pst* I or *Sph* I, respectively and for the subsequent amplification step. The use of these oligonucleotide templates involved stepwise increases in temperature (1 min intervals at 25, 30, 35 and 40 °C, then 30-second intervals at 45, 50 and 60 °C) to facilitate extension of the 3' complementary overhang to the 5' end of the non-ligated oligonucleotide templates. This created products as depicted in **Fig. 2.1 (step E)**. The resulting DNA was subsequently used as a template for PCR as described in section 2.8.

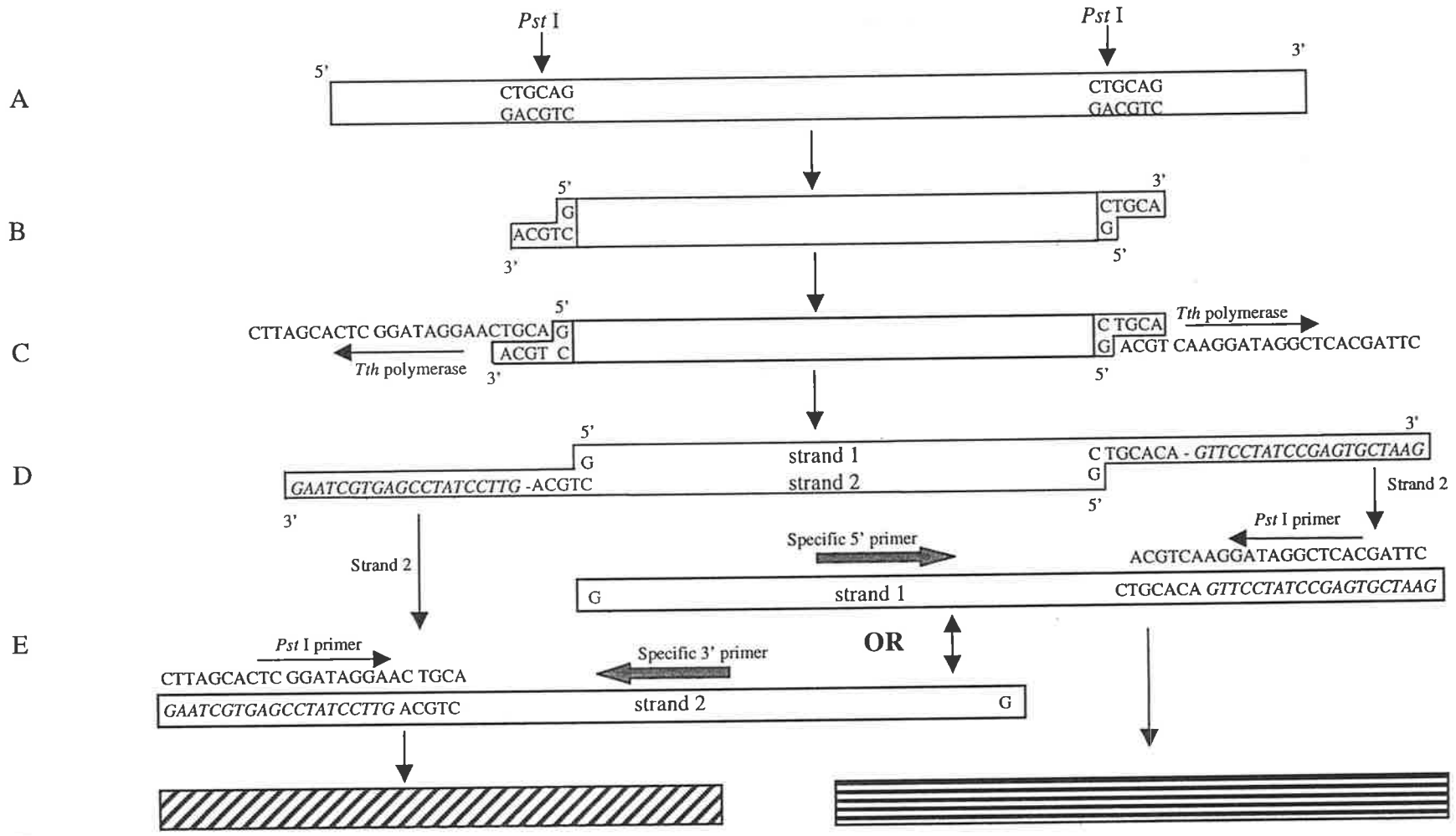


Figure 2.1. Outline of 3'-overhang-based end extension and subsequent amplification by PCR of target sequences from genomic restriction fragments.

2.8 Polymerase chain reactions (PCR)

DNA was amplified by the polymerase chain reaction using varying amounts of DNA templates in final reaction volumes of 50 µl. Reactions normally contained 50 mM KCl, 2.5 mM $MgCl_2$, 0.04% gelatin, 0.2 mM each of dATP, dGTP, dCTP and dTTP, 0.5 µM each of the 5' and 3' primers and 0.25 units of *Tth* DNA polymerase. Mixtures were overlaid with 2 drops of mineral oil. All PCR involved an initial denaturation step (95°C, 2 min), followed by 30 amplification cycles (95°C, 35 sec; 56-60°C, 45 sec; 72°C, 2-6 min) and a final 10-min extension at 72°C. Samples (10 µl) of each amplification mixture were analysed by electrophoresis (section 2.7.1).

2.9 Analysis of RNA

Reverse transcription of mRNA and analysis by PCR

Total RNA was extracted from log phase *Giardia* trophozoites and purified using QIAGEN RNeasy Spin columns (Cat. No.74103). Residual DNA was removed by incubation with RNase-free DNase. First-strand complementary (cDNA) synthesis was achieved by reverse transcription (RT) of messenger RNA (mRNA), usually using oligo-(dT) (**Table 2.1**) as a primer for polyadenylated RNA. Total RNA (2 µg in 2 µl) was mixed with oligo-dT (2 µM) and 33 units of human placental RNase inhibitor (RNA*guard*; Pharmacia Biotechnology) and water (final volume, 12 µl) and heated at 70°C for 10 min. After cooling the samples on ice, reaction buffer (10 mM DTT, 1 mM of each dNTP, 0.1 M Tris-HCl, 0.1 M KCl, 10 mM $MgCl_2$, pH 8.3) and 200 units of Superscript II reverse transcriptase (GIBCO BRL) were added to yielded a 20 µl final reaction volume. This was incubated in a thermal cycler (Corbett Research, FTS-320) for 60 min at 42 °C, then at 95°C for 10 min before cooling to 4°C. To test for the absence of genomic DNA, oligo-dT was omitted from parallel

control reaction mixtures. These yielded products in subsequent PCR when contaminating genomic DNA was present.

2.10 Cloning DNA fragments

2.10.1 Ligation conditions

Ligation of DNA was carried out in 10 µl reactions which contained 20-50 ng of linearised vector DNA, varying amounts of the restricted DNA fragment (a 3-5 molar excess of insert to vector DNA was usually used) and 1 Weiss unit of T4 DNA ligase in ligation buffer (50 mM Tris-HCl, 10 mM $MgCl_2$, 1 mM DTT, 1 mM ATP, pH 7.4). Ligation mixtures were incubated at 16°C overnight and then used for bacterial transformation. A ligation control, containing vector but no insert, was always tested in parallel transformation reactions to determine levels of undigested or recircularised vector DNA.

2.10.2 Generation of blunt-ended DNA

The following reagents were added directly to PCR tubes (final reaction vol = 50 µl) to repair the products: T4 DNA polymerase (1 unit); dNTP's (each 4 mM final); T4 polynucleotide kinase (5 units); and ATP (1 mM final). The tubes were incubated at 25°C for 25 min before the reactions were stopped by adding 3 µl of 0.5 M EDTA, (pH 8.0), mixing and then heating at 70 °C for 10 min to inactivate the enzymes. The DNA was precipitated by adding 0.1 vol of 3 M sodium acetate, (pH 5.6) and 2.5 vol of absolute ethanol, then incubated at -20°C overnight. Following centrifugation, the DNA pellets were washed with 70% ethanol, dried and redissolved in 20 µl aliquots of water. Two microliter samples were used for ligation to blunt ended, linearised cloning vectors.

2.10.3 Bacterial transformation

Competent *E. coli* (strain DH5αF') were prepared by treatment with $CaCl_2$ and RbCl (Sambrook *et al.* 1989) and stored at -70°C. To a suspension of 50-100 µl of competent cells

(thawed on ice) was added 10 µl of ligation mixture (section 2.10.1). The tube was left on ice for 30 min. The cells were then heat-shocked (90 sec, 42°C) and chilled on ice for a further 5 min. After warming to room temperature (5 min), 1 ml of nutrient broth was added and the suspension was incubated at 37°C for 45 min to allow synthesis of β-lactamase (conferring ampicillin resistance to transformed cells). The cells were then pelleted by centrifugation (800×g, 5 min) and resuspended in 200 µl of physiological saline. Aliquots (50 µl) of different dilutions were spread onto plates of nutrient agar containing 50 µg/ml ampicillin, 24 µg/ml isopropyl-β-D-thiogalactopyranoside and 25 µg/ml 5-bromo-4-chloro-3-indolyl-β-D-galactoside to enable the identification of colonies containing plasmid inserts. The cells were grown overnight at 37°C.

2.11 DNA sequencing and analysis

2.11.1 Nucleotide sequence determinations using dye terminators

DNA for sequencing reactions was purified on QIAGEN Tip-20 Micro columns. To 1 µg of DNA template were added 8 µl of dye terminator mix, 3.2 pmol of primer and distilled water to a final volume of 20 µl. Reactions were overlayered with two drops of mineral oil and placed in a thermal cycler. Thermal cycling was commenced with an initial denaturation step (96°C, 30 sec), followed by 25 cycles (each comprising a 15-sec annealing step at 50°C and a 4 min extension at 60°C). The labelled extension products were precipitated as described in section 2.7.7 .

Nucleotide sequence determinations were carried out by the IMVS Molecular Pathology Unit, using an Applied Biosystems model 373A automated DNA sequencer.

2.11.2 Analysis of nucleotide and amino acid sequences

Nucleotide and deduced amino acid sequences were analysed manually and with the aid of specialised computer programs. Nucleotide sequences were checked manually against

the chromatograms for errors or inconsistencies. Those expected to overlap were subjected to alignment using FASTA (Altschul *et al.* 1997). These sequences, derived from both strands, were compiled into larger, contiguous segments using the editing functions of DNASIS (version 7.0, Hitachi Software Engineering, San Burno, CA.). The assembled sequences were inspected for open reading frames, translated using DNASIS and submitted to GenBank. Deduced amino acid sequences were further analysed using PROSIS. Amino acid sequence alignments and homology searches were carried out locally using the FASTA feature of PROSIS. Additional similarity searches of both nucleotide and amino acid sequence databases were performed using BLAST (Altschul *et al.* 1997), at GenBank, or locally using a down loaded version of BLAST to search various updates of the *Giardia* Genome database. Multiple sequence alignments were obtained using CLUSTAL W (Thompson *et al.* 1994). These were edited manually and subjected to phylogenetic analysis using the Molecular Evolution Genetic Analysis Package, MEGA (Kumar *et al.* 1994).

2.12 Southern hybridisations

2.12.1 Preparation of DIG-11-dUTP-labelled single-stranded probes

Single stranded DNA probes labelled with digoxygenin (DIG) were synthesised by single-primer PCR using reaction mixtures containing DIG-11-dUTP. Purified template DNA (200 ng), DIG-dNTP mix (2 mM each of dATP, dCTP and dGTP, 1.3 mM of dTTP, 0.7 mM DIG-11-dUTP), 10 mM Tris-HCl, 50 mM KCl, 2.5 mM $MgCl_2$, 0.5 µM of the specific-primer and 0.25 units of *Tth* DNA polymerase were mixed, in a final reaction volume of 25 µl and overlayered with mineral oil. The DNA was denatured at 95°C for 1 min then subjected to 35 reaction cycles (95°C, 35 sec; 56-60°C, 35 sec; 72°C, 1-4 min) and a 10-min final extension at 72°C. Labelled probes were stored in -20°C.

2.12.2 Southern transfers

Following electrophoresis, staining, visualisation and photography of resolved DNA bands, the agarose gels were rinsed in water and immersed, with constant rocking, in 0.25 M HCl for 10 min. After two rinses in water, the gels were immersed in 0.5 M NaOH, 1.5 M NaCl and rocked for 25 min. DNA was transferred overnight to positively charged nylon membranes (Hybond N^+Amersham), by capillary transfer, using 0.4 M NaOH as the transfer medium. The alkali conditions denature the DNA and cause covalent attachment to the membrane.

2.12.3 Hybridisation, washing and staining of blots

Following transfer, membranes were washed twice in 50 ml of 2× PE (phosphate-EDTA) supplemented with 0.1% SDS (10 min per wash, with rocking), then incubated at 65°C for 3 hours in 50 ml of 5× PE + 1% SDS. Prior to the addition of DIG-labelled probes, membranes were incubated for 4 hrs at 42°C in pre-hybridisation solution consisting of 50% (v/v) deionised formamide, 40 mM sodium phosphate, 5× Denhardt's solution, 250 µg/ml herring sperm DNA (Sigma) and 5× SSC. This fluid was removed and replaced with fresh hybridisation buffer. Immediately before use, the labelled probe was denatured by heating at 95°C for 10 mins and added to the hybridisation buffer. Hybridisation was allowed to occur over 16 hrs at 42 °C. The membranes were subsequently given two 5-min washes, with shaking, at room temperature in 2× SSC / 0.1% SDS. These were followed by two additional 15-min washes at 42°C in 0.1%× SSC / 0.1% SDS. The membranes were finally washed in buffer 1 (0.15 M NaCl, 0.1 M Tris-HCl, pH 7.5) and blocked for 30 min in skim milk (5% in buffer 1). Alkaline phosphatase-conjugated anti-DIG Fab fragments, diluted in Enzyme Diluent (1:5000) were then added and after a 30-min incubation at room temperature, the membranes were given four 5-min washes in buffer 1 and a 2-min wash in staining buffer (0.1 M Tris-HCl 25 mM diethanolamine-HCl, 0.1 M NaCl, 2 mM $MgCl_2$, 1 µM $ZnCl_2$, pH

9.55). Colorimetric detection of target DNA was achieved by the addition of alkaline phosphatase staining solution and incubation in a sealed bag, at 37°C in the dark. Staining reactions were stopped by washing with water.

2.13 Generation of genomic DNA libraries

Genomic DNA libraries were prepared from *Sac* I generated restriction fragments of *G. intestinalis* DNA by ligation with linearised pBluescript II SK+ DNA (section 2.10.1) and transformation of competent DH5α *E. coli* (section 2.10.3).

2.13.1 Colony blotting

Bacterial colonies were lifted onto Hybond-N^+ filter discs, taking care to label both the discs and plates such that the blots could be aligned correctly after subsequent staining. The membranes were placed colony-side up onto sheets of Whatman filter paper that had been saturated with 0.4 M NaOH. They were left for 10-30 min, allowing the cells to lyse and the DNA to attach covalently to the membranes. The discs were then washed twice in 50 ml of 2× PE / 0.1% SDS. Cell debris was removed from the blots by wiping gently with cotton wool. The 'blots' were washed, incubated with diluted probe in hybridisation solution, washed post-hybridisation by incubating at 65°C for 3 hrs in 50 ml of 5× PE / 1% SDS and finally stained (section 2.14.1).

2.13.2 Screening genomic libraries

Bacterial colonies were replica-plated onto nutrient agar/ampicillin plates and screened by colony blotting (section 2.13.1). Colonies of interest were picked individually from the original plates, regrown overnight at 37°C on nutrient agar/ampicillin plates, and the resulting colonies were patched onto nutrient agar/ampicillin plates and re-tested for hybridisation by colony blotting. Positive colonies were grown overnight in nutrient broth (with added ampicillin, 50 μg/ml) and the plasmid DNA was isolated (section 2.6.1) for

analysis by restriction mapping (section 2.7.3) and Southern hybridisation (section 2.12.2, 2.12.3). Appropriate regions or segments of cloned inserts were sequenced.

2.14 *In situ* mRNA hybridisation

2.14.1 Slide treatment for *in situ* hybridisation

Microscopic slides (Sail brand, Cat. No: 7105) were treated with 3-aminopropyltriethoxysilane (AAS, Sigma Cat. No A-3648) to increase the adhesion of *Giardia* trophozoites to the glass surface. The slides were first cleaned by washing overnight in chromic acid (100 g of $K_2Cr_2O_7$ in 85% H_2SO_4) following by 3-4 rinses in distilled water. Dried slides were subsequently immersed in absolute ethanol (3 seconds), then in 10 ml of 3-aminopropyltriethoxysilane in 500 ml absolute ethanol (3 seconds) and rinsed successively in absolute ethanol and distilled water. To improve the adhesion of cells, the slides were subsequently soaked in 1% glutaraldehyde for 30 min, washed well in distilled water and dried. These slides stored at room temperature.

2.14.2 *In situ* mRNA hybridisation

After collection of trophozoites (section 2.5), the slides were placed in a Hybaid Omnislide slide rack at 37°C and 200 µl of a *Giardia* suspension (1×10^6 cells) in PBSm was placed on the centre of each slide. After 10 min, excess fluid was removed by flicking the slides and the cells were immediately fixed in 3% paraformaldehyde (20 min, room temperature). After rinsing 3 times in PBS, the slides were immersed in 0.1 M glycine (5min), then rinsed again in PBS before being given successive 5-min immersions in 70% ethanol, absolute ethanol, and being finally dried in a 37°C incubator. A 4 µl aliquot of hybridisation solution (50% deionised formamide, 5× SSC, 50 mM sodium phosphate (pH 6.4), 10 mM EDTA, 40 mM Tris-HCl, 10 mg/ml herring sperm DNA, 5× Denhardt's solution), containing a single stranded DIG labelled probe, was placed on to the cells. After

overlaying with a coverslip, this was sealed with nail polish and the slides were placed in a Hybaid Omnislide hybridisation cycler, heated to 80°C (5-min denaturation) and the temperature was then reduced to the chosen hybridisation temperature (as determined from the theoretical melting temperature of the probe). After a 3-hr hybridisation step, the coverslips were removed using a scalpel blade and the slides were given three 10-min washes at 65°C in 2× SSC / 0.1% SSC, followed by a blocking step in skim milk (3% in buffer 1, 30 min). One µl (1.3 units) of alkaline phosphatase-conjugated anti-DIG Fab fragments was diluted in 750 µl of Enzyme Diluent and 40 µl were placed on each slide, covered by a coverslip and left in a humid box for 30 min at room temperature. The coverslip was then removed and the slides were washed twice with buffer 1 (15 min). Staining was allowed to occur overnight in the dark at room temperature by leaving the slides in alkaline phosphatase staining solution containing 0.1 M Tris-HCl (pH 9.5), 0.1 M NaCl, 0.5 mM $MgCl_2$, 175 µg/ml 5-bromo-4-chloro-3-indolyl-phosphate and 337.5 µg/ml 4-nitro blue tetrazolium. The slides were finally rinsed in water. A 25% mixture of glycerol in PBS was used as a mounting solution and slides were viewed by bright field microscopy.

Chapter 3

The '*vsp*136' gene subfamily

The '*vsp*136' gene subfamily

3.1 Analysis of Ad-1/c3 trophozoites for vsp gene transcripts

As was mentioned earlier (section 1.18), before commencement of this project, two vsp genes, *vsp*417-2 (*tsp*11) and *vsp*52, were identified by screening an Ad-1 *G. intestinalis* genomic library for expressed antigens using an antiserum raised against the purified VSP (the 2D4 mAb-reactive surface antigen complex) of Ad-1/c3 trophozoites. This raised the question of whether the larger (~70 kDa and 80-85 kDa) subunits of the Ad-1/c3 surface antigen complex were in fact the 69-Kda and 89-Kda polypeptides encoded by these two characterised genes. At the beginning of this project, experiments were therefore carried out to test the hypothesis that *vsp417*-2 and *vsp52* are expressed by the majority of the cells in Ad-1/c3 cultures. Initially, RNA was extracted from these cells and tested for the presence of vsp gene transcripts. Experiments carried out in the first few months indicated that *vsp*52 was not expressed in the majority of Ad-1/c3 trophozoites and therefore these results are not presented in detail. However, these preliminary experiments led to the discovery of three novel vsp genes which are very similar to the *crp*136 gene described by Chen *et al.* (1995). These findings are described in detail and they constitute the bulk of this chapter.

3.2 Analysis by RT-PCR

Total RNA was extracted from *Giardia* Ad-1/c3 trophozoites and the mRNA was reverse-transcribed from the 3' polyadenylated end using primer 80, an oligo-(dT) (**Table 2.1**). The resulting cDNA was analysed by PCR for the presence of *vsp*417-2 (*tsp*11), *vsp*52 and other known vsp gene sequences.

3.2.1 Detection of *vsp*417 subfamily transcripts

Using locus-specific primers, amplified DNA corresponding in size to the expected segments of the *vsp*417-2 (*tsp*11) gene were amplified from Ad-1/c3 cDNA and the identity of the products was supported by restriction analysis (data not presented). However, in parallel reactions, the use of consensus primers that were specific for other *vsp*417 subfamily genes yielded similar amounts of products that corresponded to transcripts from additional loci, e.g. *vsp417-1* (*tsa*417) and *vsp417-6* (data not shown). This probably reflected the presence, within the bulk RNA, of trace amounts of mRNA from variants that had emerged within the culture but which were still present as minor subpopulations. The data thus provided no clear evidence that *vsp417*-2 was expressed by the majority of the cells, i.e. by the 2D4 mAb-reactive trophozoites.

3.2.2 Failure to detect transcripts of the *vsp*52 gene

The cDNA, prepared from Ad-1/c3 trophozoite mRNA by oligo-(dT)-primed reverse transcription, was also tested in PCR for the presence of transcripts from the *vsp*52 gene. Using a primer (oligo 17) that, at the time, was presumed to be specific for this gene, amplification products that were similar in size to the expected 3' segment(s) of *vsp*52 were obtained (**Fig. 3.1**, lanes c & d). However, attempts to amplify either the full *vsp*52 coding sequence, using oligos 23 + 731 or 23 + 80 that were specific for this gene, or segments from the 5' portion of the coding sequence (using oligos 23 + 21) were unsuccessful (**Fig. 3.1**, lanes a & b). The quality of the cDNA, i.e. that it represented full-length mRNA, was confirmed by the successful amplification, from the same cDNA preparations, of near full-length segments of the glutamate dehydrogenase (GDH) gene using forward primers that were specific for the 5' end of this gene (data not shown). Restriction analysis of the two PCR products obtained using oligos 17 + 731 and 17 + 80 (expected to be derived from the 3' portion of *vsp*52) showed that they were both different from the corresponding

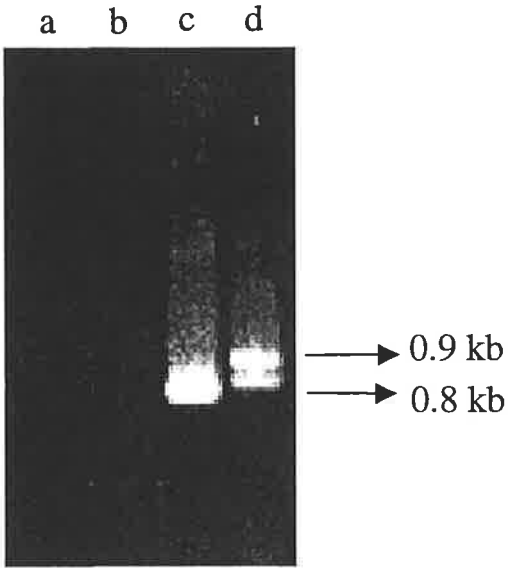


Figure 3.1. Attempts to amplify segments of the *vsp*52 gene from Ad-1/c3 cDNA by PCR. Reaction mixtures contained primer combinations designed to amplify different segments of the *vsp*52 coding sequence. (a) Oligos 23 + 80 (full coding sequence); (b) Oligos 23 + 21 (5' segment); (c) Oligos 17 + 731 and (d) Oligos 17 + 80 (3' segments).

characterised segment of the cloned *vsp*52 gene (data not shown). To identify these cDNA-derived products properly, each was cloned into pGEM-7Zf(+) and sequence data were obtained from two constructs that contained the respective ≈0.8-kb and 0.9-kb inserts. Each of the cloned fragments represented single reading frames (ORF) that appeared to be a part of longer coding sequences. Each also showed similarity with the expected 3' portion of the *vsp*52 coding sequence. The first insert, designated *vsp*R1 (**Fig. 3.2**), was very similar to another characterised vsp gene (*crp*65, Chen *et al.* 1996) with which it shared nucleotide identity at 98% of sites over an 827-bp overlap (data not presented). The second insert, designated *vsp*R2, exhibited a common primer (oligo 731) sequence at both ends, indicating that this degenerate reverse primer (corresponding to the transmembrane-encoding 3' end of vsp genes) had acted as both the forward and reverse primer in the amplification of this DNA. Interestingly, despite this apparent mis-priming, the *vsp*R2 sequence exhibited significant similarity with the *crp*136 gene (described by Chen *et al.* 1995). It possessed two copies of a tandem repeat element, one of which was truncated at the ligated 5' end of the insert (**Fig. 3.2**), and it exhibited nucleotide identity at 80% of sites over an 840-bp overlap with *crp*136 (data not presented). The amino acid sequences encoded by the characterised segments of the *vsp*R1 and *vsp*R2 ORF were aligned against all available VSP sequences. This indicated that both were closely related to VSP 52 as well as CRP136 and CRP65 (**Fig. 3.16**) and that they might all represent loci that have diverged from replicate copies of a common ancestral gene. The RT-PCR data indicated, however, that Ad-1/c3 trophozoites were not expressing the *vsp*52 gene and that the 85-90 kDa component of the surface antigen complex on these cells must be encoded by a different gene. Consequently, subsequent efforts, representing the bulk of this thesis, were directed at elucidating the nature and identity of the genomic loci represented by the *vsp*R1 and *vsp*R2 cDNA sequences and additional closely-related vsp gene loci, in order to gain an insight into the evolution of these particular genes and the structural differences between the encoded polypeptides.

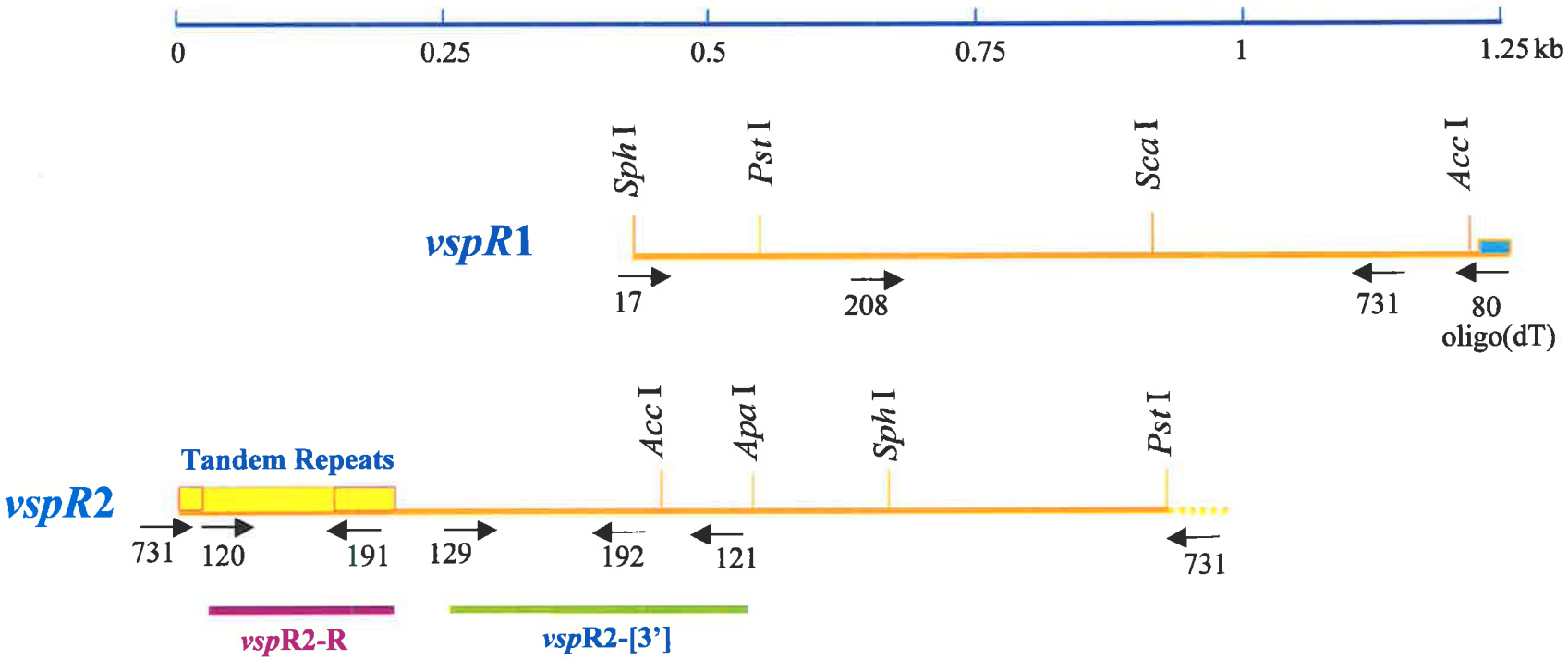


Figure 3.2. Segments of two VSP gene transcripts, *vsp*R1 and *vsp*R2, identified by RT-PCR in RNA extracted from Ad-1/c3 *G. intestinalis* trophozoites. For *vsp*R1, the blue box represents the poly-A tail of the mRNA (identified by sequence analysis). For *vsp*R2, yellow boxes represent tandem repeat elements (one full and two half units, identified by nucleotide sequence analysis). Bold purple and red lines represent segments amplified using oligos 120 + 191 (repeat segment, '*vsp*R2-R') or 129 + 121 (non-repeat segment, '*vsp*R2-[3']) respectively. Diagnostic restriction sites are indicated. PCR primers sites are depicted as arrows. Segments of *vsp*R2 that were used as probes for hybridisation analyses are represented by bold lines (*vsp*R2-R = repeat region; *vsp*R2-[3'] = non-repeat downstream region). The sequences are available under the GeneBank accession numbers as indicated:*vsp*R1, AF298867; *vsp*R2, AF300879.

3.3 Identification of related vsp genes in the genome

In situ mRNA hybridisation of Ad-1/c3 trophozoites (98% of which were labelled with mAb 2D4), using anti-sense probes derived from different segments of the cloned *vsp*52 gene, showed that only a very small minority (<0.05%) of the cells contained transcripts from this locus (data not presented). This confirmed the RT-PCR results, which had failed to detect *vsp*52 mRNA in bulk RNA extracted from these cells (section 3.2.2). However, with evidence of related transcripts (*vsp*R1, *vsp*R2) within this clonal line, a decision was made to investigate the number and nature of these *vsp*52/*vsp*136-like loci within the genome.

To determine how many distinct sequences related to *vsp*R2 exist within the genome, Southern hybridisations were performed on *Pst* I and *Sac* I-restricted Ad-1/c3 chromosomal DNA. Two segments of the cloned *vsp*R2 cDNA were used to prepare DIG-labelled probes for hybridisation analysis. The first probe, designated *vsp*R2-R and amplified using oligos 120 + 191, comprised the 120-bp repeat element (**Fig. 3.2**). The second probe, designated *vsp*R2-[3'], comprised a 275-bp segment that was amplified by oligos 129 + 121 and corresponded to the 3' non-repeat portion of the cloned insert (**Fig. 3.2**). The latter probe was almost identical to a segment in the 3' portion of *crp*136. The results of these hybridisations are shown in (**Fig. 3.3**). Significant differences were observed in the number and size of genomic restriction fragments that hybridised with the two *vsp*R2-derived probes. At least 10 *Pst* I and *Sac* I fragments could be seen to hybridise with variable strength to the *vsp*R2-[3'] probe (**Fig. 3.3A**). In contrast, only 3-4 fragments hybridised strongly to the *vsp*R2-R probe, although the upper *Pst* I band may have represented multiple fragments and 3-4 additional weakly-hybridising fragments were evident (**Fig. 3.3B**). The size range of the fragments detected by the two probes was different. In the case of the *vsp*R2-R probe, the largest *Sac* I and *Pst* I fragments were approximately 8.6 and 3.7 kb respectively, whilst the largest fragments detected with the *vsp*R2-[3'] probe were much longer (17-19 kb; **Fig. 3.3A**). The bands evident in **Fig. 3.3** represented numerous genomic sequences that were related to

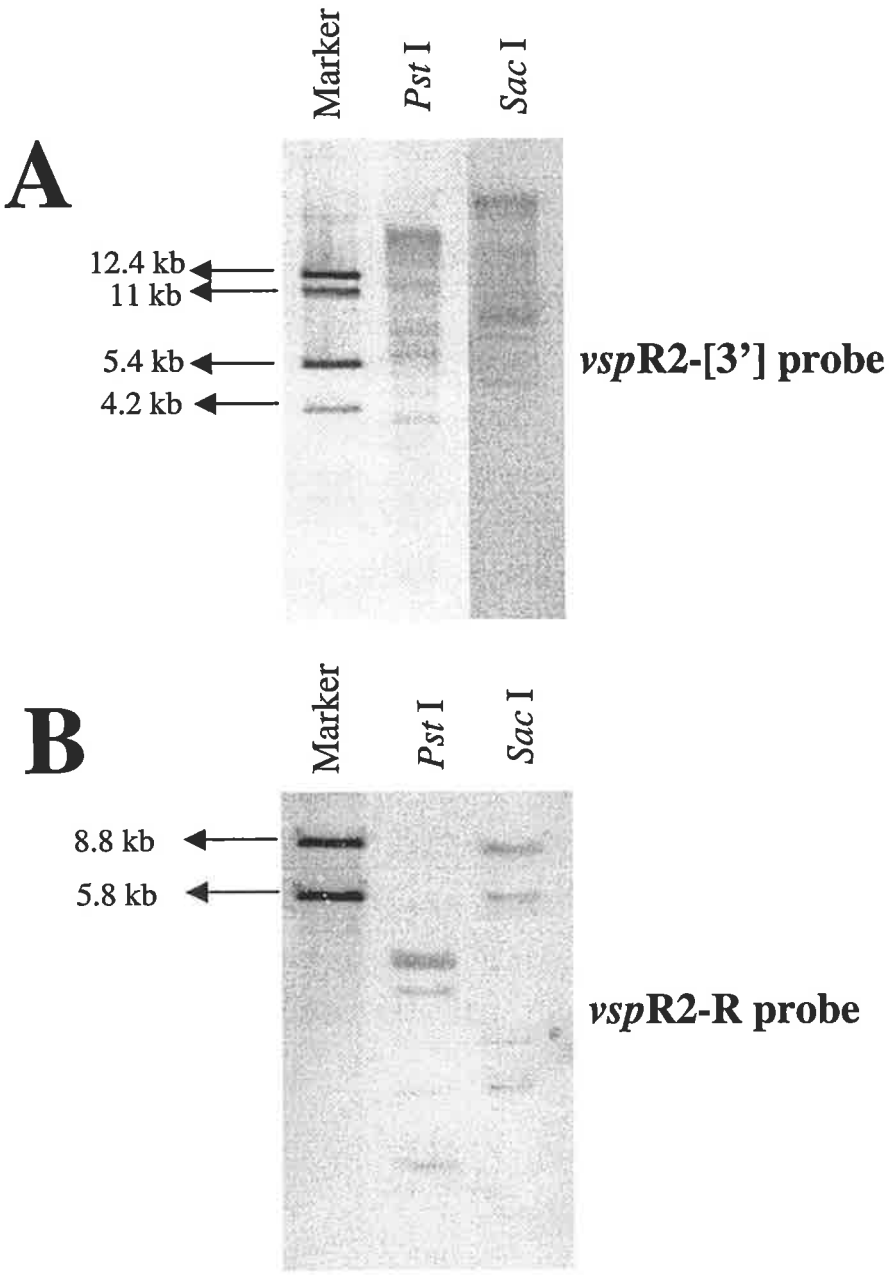


Figure 3.3. Southern hybridisation analysis of restricted *G. intestinalis* chromosomal DNA. Replicate samples of DNA were restricted by incubation with the indicated endonucleases. The fragments were separated by electrophoresis and tested for hybridisation with DIG-labelled probes. The sizes of the markers are indicated.

*vsp*R2, but it was unknown whether these represented intact vsp genes or whether they contained identical or different (but similar) tandem repeat elements.

3.4 *Giardia* genomic DNA library construction and screening

At the time these experiments were commenced, complete sequence data were available for only four giardial vsp genes (*vsp*A6, *vsp*C5, *crp*136 and *crp*65) that encoded VSP possessing tandem repeat elements (Adam *et al.* 1988b, 1992; Yang *et al.* 1994; Chen *et al.* 1995, 1996), although alleles of *vsp*A6 (*vsp*A6.2, *vsp*A6.3) and *vsp*C5 (vspC5-S1, *vsp*C5-S2) had been partially characterised (Yang & Adam 1994, 1995; Yang *et al.* 1994). The aim of the work described in this chapter, therefore, was to isolate, identify and characterise the complete *vsp*R2 gene and any closely related loci whose presence in the *Giardia* genome could be detected by Southern hybridisation, as shown above. The strategy used to obtain these loci was to construct *Giardia* genomic DNA libraries and screen colony blots for hybridisation with the *vsp*R2 probes. Three libraries were constructed in pBluescript II SK(+) from *Sac* I-restricted Ad-1/c3 genomic DNA. One of the plates which yielded a clone of interest with the *vsp*R2-R probe is shown in **Fig. 3.4**.

Five clones of interest were identified by screening a total of 100 plates, representing colonies derived from three independently-constructed *Sac* I genomic libraries. Two of these colonies yielded plasmid constructs pM2-6 and pM6-1, which contained inserts of a similar size (≈ 5.8 kb). Another two colonies harboured constructs pM42-2 and pM49-1, which also contained inserts of a similar size (≈ 8 kb). Characterisation of the latter two plasmids by restriction mapping, Southern hybridisation and partial sequence analysis showed that they contained copies of the same genomic fragment. The fifth colony which stained more faintly in colony hybridisation than the positive control colony (containing the *vsp*R2 plasmid), yielded a plasmid construct, pM24-1, containing a *Sac* I genomic insert of ≈ 3.4 kb.

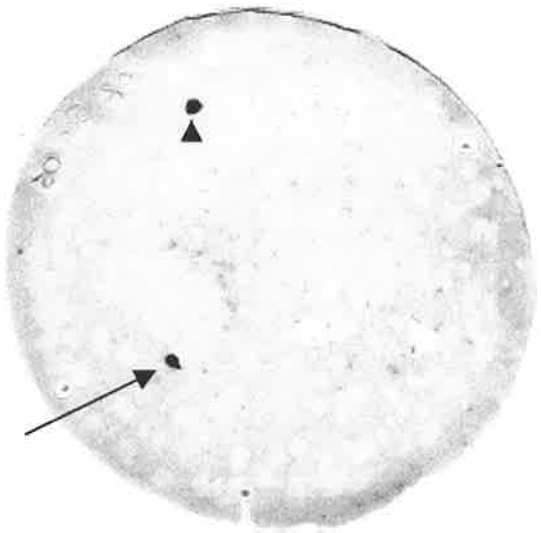


Figure 3.4. Detection of a clone harbouring plasmid pM2-6 by colony blot hybridisation. Colonies representing part of a genomic library generated from *Sac* I fragments of *G. intestinalis* DNA. The colonies were screened for hybridisation with the DIG-labelled ***vsp*R2-R probe**. One colony (arrowed) contained plasmid DNA that hybridised with the probe. The triangle (top) shows the *vsp*R2 (positive control) clone. The plate contained a total of approximately 100-300 colonies.

3.5 Characterisation of the pM6-1 insert

Restriction enzyme (e.g. **Fig. 3.5**) and Southern hybridisation analysis of pM2-6 and pM6-1 indicated that they contained the same *Sac* I genomic fragment which had been cloned in opposite orientations within the two constructs. Bands representing common fragments were evident among the restriction endonuclease cleavage products, as exemplified in **Fig. 3.5**. Both plasmids were examined by Southern hybridisation analysis (e.g. **Fig. 3.6**) in order to identify which restriction fragments hybridised with the *vsp*R2-R probe. This information was used to subsequently subclone the related fragments. The pM6-1 construct was chosen for more detailed analysis and five subclones were constructed by utilising restriction sites identified within the insert and those within the multiple cloning site of the vector. These were used to determine the nucleotide sequence of the entire 5.75-kb insert of the parent (pM6-1) construct, as depicted in **Fig. 3.7**. Analysis of the compiled sequence revealed only a single ORF of 3,825 bp, which was designated *vsp*136-2. The ORF consisted of three distinct regions (**Figs. 3.7, 3.8**). The first region comprised a 69-bp 5' segment of which the first 39 bp encoded a 13-residue hydrophobic segment (a typical signal peptide). This was predicted (using the 'Signal P' V2.0 Web Server) to be cleaved on the N-terminal side of Ala^{14} (**Fig. 3.8**). This 5' segment was followed by a second region comprising 23.5 tandem copies of a 120-bp repeat element. The central portion of this region was not sequenced. However, the number of tandem copies was calculated from the sizes of various restriction fragments that encompassed this region. The sizes of the fragments were determined as accurately as possible, using linear regression analysis of the electrophoretic mobilities and known marker fragments as size standards. The number of repeats was then calculated using knowledge of the nucleotide sequence at both ends of the repeat region (this included the first 2 to 3 tandem copies at each end). The third region of the *vsp*136-2 ORF was defined by a 933-bp segment that lacked repeat elements and formed the 3' end of the gene. This included a putative polyadenylation signal sequence (AGTAAA) that was situated 21-bp beyond the stop codon

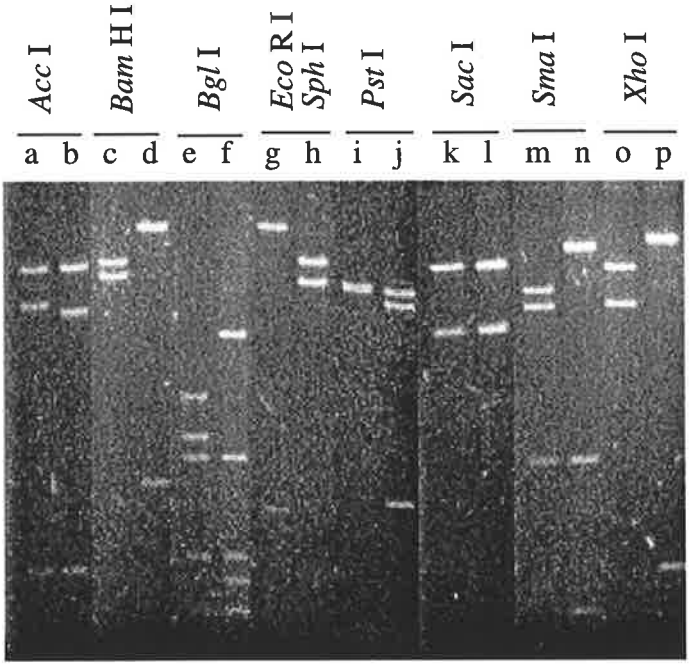


Figure 3.5. Comparative restriction analysis of plasmid constructs pM2-6 and pM6-1. Replicate aliquots of plasmid DNA were incubated with different restriction endonucleases as indicated. The reaction mixtures were subjected to electrophoresis on 1% agarose and stained with ethidium bromide. For some enzyme digests, one or more bands representing common fragments derived from the two plasmids can be seen.

pM2-6 Lanes a, c, e, g, i, k, m, o

pM6-1 Lanes b, d, f, h, j, l, n, p

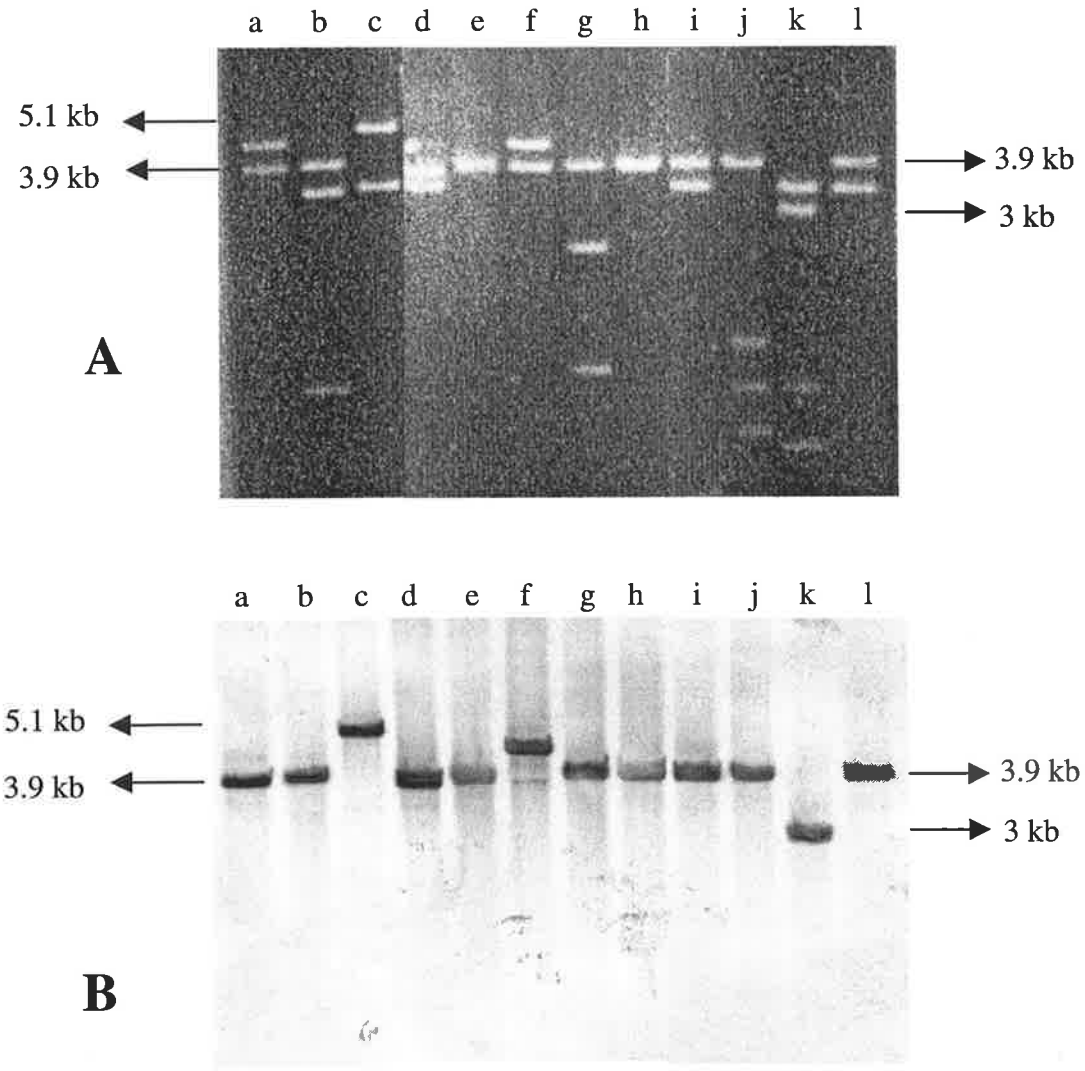


Figure 3.6. Single- and double-enzyme restriction and Southern hybridisation analyses of the pM2-6 construct.
Replicate aliquots of plasmid DNA were incubated with different restriction endonucleases. The reaction mixtures were subjected to electrophoresis in 1% agarose and stained with ethidium bromide (A). Gels were subsequently analysed by Southern transfer and hybridisation (B) using the DIG-labelled ***vsp*R2-R probe**. Size markers are indicated. The enzymes used were :

a	*Bgl* II	g	*Bgl* II + *Dra* I
b	*Sma* I	h	*Bgl* II + *Bam* H I
c	*Xho* I	i	*Bgl* II + *Xho* I
d	*Pst* I + *Xho* I	j	*Sma* I + *Dra* I
e	*Pst* I + *Sac* I	k	*Sma* I + *Sph* I
f	*Bam* H I	l	*Sma* I + *Bam* H I

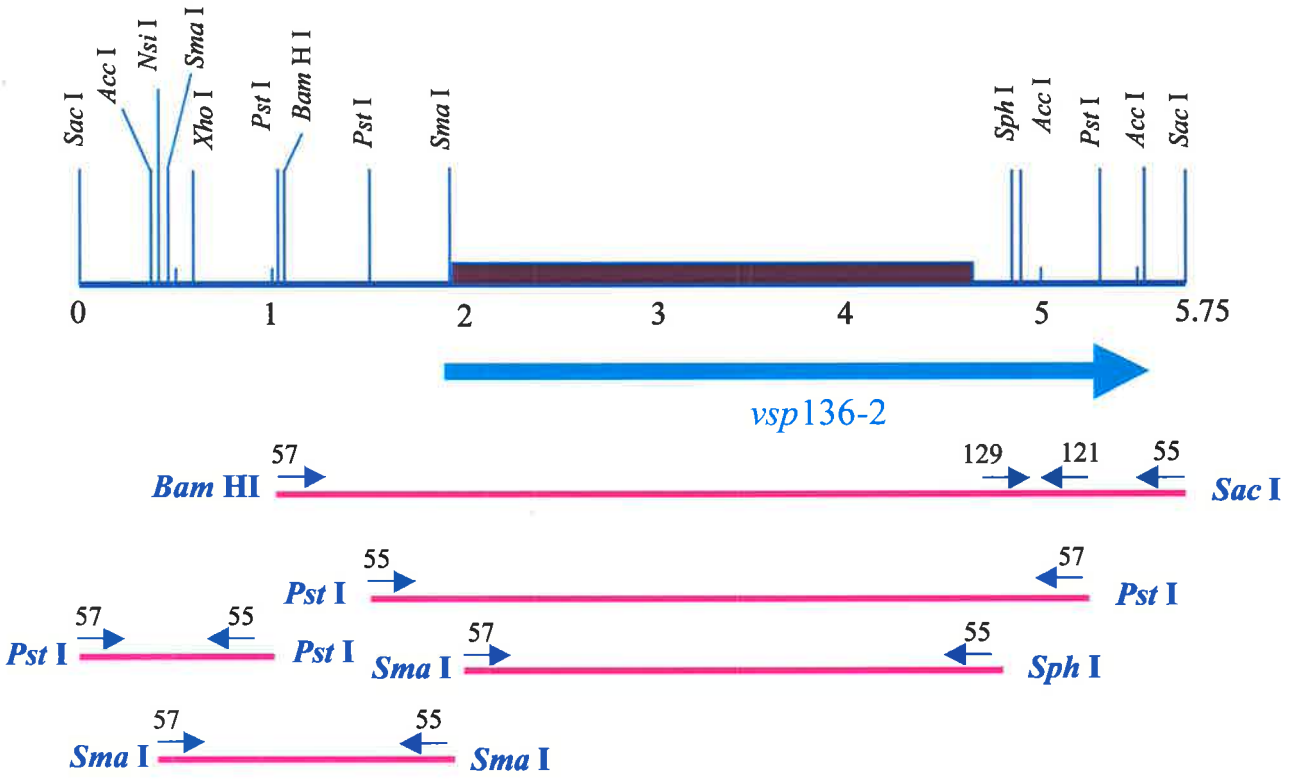


Figure 3.7. Schematic representation of the pM6-1 insert. The single open reading frame, designated *vsp*136-2 is denoted by a thick arrow. The tandem repeat region within this open reading frame is represented on the scaled line by a purple shaded box. Subcloned fragments used for sequencing are depicted by bold lines. Restriction sites that were identified experimentally by cleavage and confirmed by sequence analysis are shown. The identity and annealing positions of the primers that were used for sequencing are shown by arrowheads.

Figure 3.8. Nucleotide and deduced amino acid sequences specified by the single open reading frame identified within the 5,755-bp insert of pM6-1. The 3,825-bp coding sequence (*vsp*136-2) is indicated. The deduced N-terminal signal peptide and C-terminal invariant segment, 'CRGKA', are underlined, as also is the stop codon, polyadenylation signal sequence (AGTAAA, double underlined) and VSP gene-specific pre-poly(A) signal sequence (CTTAGRT). The sequence has been submitted to Genbank and is available under accession number AF249878.

```
       Sac I
   1   GAGCTCGTACGGGGTGAGGCGGCGCCGGTGCTCGCCGGCCTTCACGTCGGCCTCGCCGGGGGGCGGATG
  70   CTGGGCAGCCCCTGCGACCGGCCCCAGCTGCCCGTCTGCGTAGATGTCCCGCCCGTCGCGCGCGGGGAC
 139   GGCGACCGTGCGCCTCGTGAAGGACGCCTCGATGACCACCGTGCTCTCTGTGCGCACCGCAGGCATCCG
 208   GGCCTCGGGGTCGGCGGGGGGCCTCGGCGCCTCGTCCGGGGCCTCCTCCACCTGCCGCCGCAGAGCCGA
 277   GACCTGCTGCTCGCGCATCCTCCGCTCCTCGCCGGCCTGCCGCTGGCGGGCCTCGATGTCCGCGTCGAC
 346   CTGCCTGCGCTTCTCGGCCGCCGCGGCCTCTGCCTCCCTCCTGGCCTGCTCCCGGACGTATGCATCATA
 415   CCGCGCATAGGACTCCTCGCGGCGCCTCGCGGCCTCCTCGGGGCTCCCGTCGAACTCCAGCGCAGCCCA
 484   CGTGCCCGGGGTCTTCTTCAGGAGGTGGACCACGAGGCATCCGTCGCTGAACTTGCAGCGCGCGTTCTC
 553   CAGCTGGACGGCGTCGTAGAGGTCTATGTTGGAGACGTAGCGCTCCGGCGCAGAGATCGTGATGCACAG
 622   GTCGGTGACCGTCACCTCGAGCGACCGCGGCGGCACGCCCGCCGGGACCCGCACGCAGGCCGTGAGGGC
 691   CTCGGGGGTCTGCGACAGGGTGAACTGGAACGAAAGAGCCATCCGGGGCGCGGGCAAGCGAGGCGGGGT
 760   TAAATGAAAATAGGCCGCAGCGTGGGGCTGCCCTGTGCGAGACGGCTCCTTCGCGTCTGGCGTTCTGTC
 829   CAGAGCTATTCTGTACGTCCCAGCCATGCGGTAGGGAGACACCCGATCAGCTAGCTCACCATGGTTGAC
 898   GAGCCCCGGCACTGGCCACCTACCCAGTCTGTGGAGGGCCACGGTCCACCGACGCCTCCTGCTCGGCAC
 967   ACGCTTGCCGTCCAGAGAGCCAGAGGATGGCCCGGCTGTCCCTATGTGGAGCACTGCAGTCGCGGAGTC
1036   AGCCGCCGGCGCTCTCCCCCTCATAGGGCGGAGCTGCATCTGTTACCTCCGGGGATGGATCCAGGTGGT
1105   CACGGGGCGCCGAGACACTGTCTGGCGGTGCACCTGAGCACGTCTCTCCGCTGAAGCCTTTGGCTCTGC
1174   AATGGGGTCTGGGGCGAGTAGGCCCACACGTCAGCGGACCCTGCGGCCAGGCAGCACAGGAGGCTGTGT
1243   ACTCCAATATGGTGCGCTGGACGAAGGAGCACGAGGCCGGTCCGGGGGAAGAGAGGCGTGGATGGAAAG
1312   CAACAGGCTCGTGGTGGTCAGATGAGAAGCGAAAGTTATTGGGGTCTATCTTCTGCTCCCTTGTGACCA
1381   GTGAGAGCATTCGCTTAGAGATCTCTAATCAGCAGCTTTATAGAATAGCGCCGCAGAAGCCATGGAGCG
1450   AGCAGATGTAATGTGACGCATTTACTACCGGGACGCCCTGCCCCTGCAGAGGAAGCCAACGAGCACAGC
1519   GACAGCAATGTACAGGAGTGCACCCCGAACAATCATTCCACAAGAGAGGTGCGATCTGAGTCTGGCCCT
1588   CACCGATCCTGTTCCTATCGAAGTCAAGGGGAACTGCACTATCAACTAGCCTCCGATGGGTGTGCTCCT
1657   ATCTCGTCCTCCTCCAGTCCAACTGATCAAGAGAGCAAGAGCGGTTGCCTCTCACTGCTCGCGGCAACA

VSP136-2:                                                  M  L  V  G  F  L  L  V  C
1726   CATCAGCCTGATGAAAAAGCAGTAGCACACGAACTCCGGCTAATGTTGGTAGGATTTCTCCTAGTTTGT
  10   A  T  V  L  A  K  D  S  G  K  D  T  C  F
1795   GCAACCGTGCTGGCAAAGGACAGCGGGAAGGACACATGCTTC

Repeat region (23 copies):
  24   P  G  Y  T  L  N  T  D  T  K  Q  C  T  K  D  P  E  A  P  C  N  V  E
1837   CCGGGCTACACCCTCAACACTGACACGAAGCAGTGCACCAAGGACCCAGAAGCGCCGTGCAACGTCGAG
  47   G  C  E  T  C  V  E  G  N  A  Q  Q  C  K  T  C  R
1906   GGCTGTGAGACCTGCGTGGAGGGCAACGCCCAGCAGTGCAAGACGTGCCGT

+ One half-copy:
 944   P  G  Y  T  L  N  T  D  T  K  Q  C  T  K  D  P  E  A  P  C  N
4597   CCGGGCTACACCATCAACACTGACACGAAGCAGTGCACCAAGGACCCAGAAGCGCCGTGCAAT

Non-repeat 3' segment:
 965                        T  P  N  C  K  T  C  D  N  P  K  T  D  N  E  I
4660                        ACTCCCAACTGTAAGACCTGTGACAACCCCAAAACAGACAACGAAATC
 981   C  T  K  C  N  D  G  D  Y  L  T  P  T  N  Q  C  V  P  D  C  T  A  I
4708   TGCACTAAATGTAATGATGGCGACTACCTCACCCCTACAAACCAATGCGTACCTGACTGCACAGCCATC
1004   S  G  Y  Y  G  D  T  D  K  K  C  K  A  C  N  P  E  C  A  E  C  V  G
4777   AGCGGATACTACGGAGATACTGACAAGAAGTGTAAGGCATGCAACCCTGAGTGCGCTGAGTGCGTTGGA
1027   P  A  N  N  Q  C  T  A  C  P  V  G  K  M  L  Q  Y  T  D  T  N  T  P
4846   CCGGCCAACAATCAGTGCACGGCCTGTCCTGTTGGGAAAATGCTTCAGTATACAGATACTAATACTCCT
1050   V  N  G  G  T  C  M  D  Q  C  S  V  S  S  T  N  D  G  C  A  E  C  G
4915   GTCAATGGGGGCACGTGCATGGACCAGTGCAGCGTGAGTTCTACAAACGATGGATGCGCAGAGTGTGGG
```

Figure 3.8 (pM6-1), p. 2

```
1073  A  Q  I  G  G  T  A  Y  C  S  K  C  K  N  T  Q  Q  A  P  L  N  G  N
4984  GCTCAGATAGGAGGAACTGCATATTGCTCAAAGTGCAAGAACACTCAACAAGCCCCCTTGAACGGCAAC
1096  C  A  A  S  S  R  V  A  F  C  A  T  I  T  S  G  A  C  T  K  C  N  E
5053  TGTGCGGCCAGTTCACGAGTAGCCTTCTGTGCAACAATCACAAGTGGGGCCTGTACAAAATGCAATGAG
1119  G  Y  F  L  K  D  G  G  C  Y  Q  T  D  R  Q  P  G  K  Q  V  C  S  N
5122  GGTTACTTCCTCAAGGACGGCGGCTGCTATCAGACAGATAGACAGCCTGGTAAGCAGGTGTGTAGTAAT
1142  A  Q  G  G  N  G  K  C  Q  T  C  A  N  G  L  A  A  S  D  G  N  C  A
5191  GCACAGGGAGGCAATGGTAAGTGTCAGACATGTGCCAATGGCTTAGCAGCAAGTGATGGCAACTGTGCA
1165  E  C  H  S  T  C  A  T  C  S  T  A  D  A  A  D  K  C  K  T  C  A  T
5260  GAATGTCATTCTACTTGTGCTACGTGCTCGACTGCAGATGCGGCTGACAAGTGCAAGACCTGTGCCACT
1188  G  Y  N  K  E  N  G  D  D  T  T  A  G  L  C  K  K  C  S  E  K  I  S
5329  GGGTATAACAAGGAAAACGGTGACGATACCACTGCTGGGTTATGCAAGAAGTGCTCAGAGAAGATCTCT
1211  G  C  K  Q  C  V  S  S  S  G  S  S  V  I  C  L  E  S  E  V  G  T  G
5398  GGATGCAAGCAATGTGTGTCGTCATCTGGCAGTTCAGTCATATGCTTAGAATCAGAAGTAGGCACTGGT
1234  G  S  V  N  K  S  G  L  S  T  G  A  I  A  G  I  S  V  A  V  I  V  I
5467  GGAAGCGTCAACAAGAGCGGCCTCAGCACGGGGGCCATAGCCGGTATCTCTGTTGCTGTGATTGTTATT

1257  V  G  G  L  V  G  F  L  C  W  W  F  L  C  R  G  K  A  *
5536  GTAGGTGGTCTTGTCGGATTCCTTTGCTGGTGGTTCCTCTGCCGCGGGAAGGCGTAGACTTAGCTGTGT

5605  ACTTAGGTAGTAAACGCGTTACTTTATGTAGCTCGTGTAGATGTGCTGCTGGAGCTATCTGAGCGCGAT
5674  CCAGAGATCATGCGGGTAATCCTCGCTCTTTCCTGGCCGTCCGGAAACGATGGCTGGTCGTGGATAATG
5743  TAACGAGGAGCTC
            Sac I
```

and preceded by the vsp gene-specific motif, 'CTTAGGT' (**Fig. 3.8**). Overall, these features represent an intact, presumably functional gene that encodes a VSP consisting of 1,274 amino acids. Comparisons of the *vsp*136-2 coding sequence (and its encoded polypeptide) with the nucleotide (and encoded amino acid) sequences of other vsp genes revealed that it was nearly identical (99.9% identity) over its entire length with the *crp*136 gene of Chen *et al.* (1995). This ORF was therefore designated *vsp*136-2 (**Figs. 3.7, 3.8**). The sequence similarity of *vsp*136-2 with *crp*136 extended beyond the coding sequence into both the 3'- and 5'-untranslated regions, but in the 5' untranslated region it extended only 238 bp upstream from the start codon (**Fig. 3.18**). Between this point and the start codon, the two genes shared 99.2% nt sequence identity (only 2 nt differences). This is discussed further in section 3.11.

3.6 Characterisation of the pM24-1 insert

As mentioned earlier (section 3.4), the pM24-1 construct hybridised only weakly with the *vsp*R2-R probe in comparison to the hybridisation intensity observed with some of the other constructs, e.g. pM6-1 and pM42-2. **Figure 3.9** shows the difference in hybridisation staining intensity observed for similar molar amounts of fragments from the pM24-1 and pM42-2 plasmid constructs.

The different staining intensity observed for these two cloned inserts was explicable on the basis of two obvious possibilities:- differences in the number of tandem repeat elements in each insert (if both inserts contained an ORF with same repeat unit), or a significantly lower level of nt sequence identity between the pM24-1 insert and the *vsp*R2-R probe.

To address this problem, both ends of the pM24-1 insert were sequenced using T3 and T7 primers and on the basis of the sequences obtained, six additional primers were designed to generate overlapping sequences by 'primer walking'. A relatively short, 1.1-kb

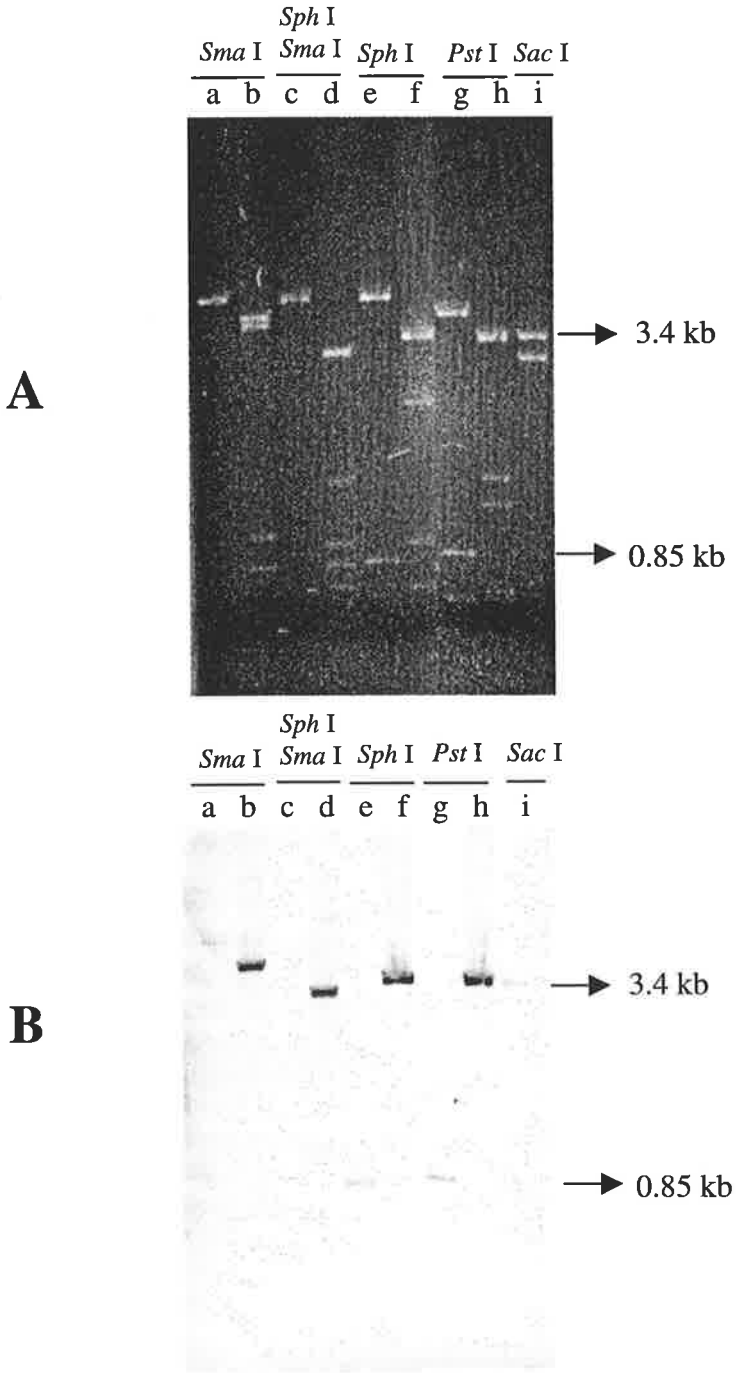


Figure 3.9. Analysis of restriction fragment length polymorphisims between the pM24-1 and pM42-2 inserts. Restriction fragments resulting from cleavage by the indicated endonucleases were separated electrophoretically and stained (A) with ethidium bromide or (B) after Southern hybridisation with DIG-labelled ***vsp*136-R** probe.

pM24-1: Lanes a, c, e, g and i

pM42-2 Lanes b, d, f and h

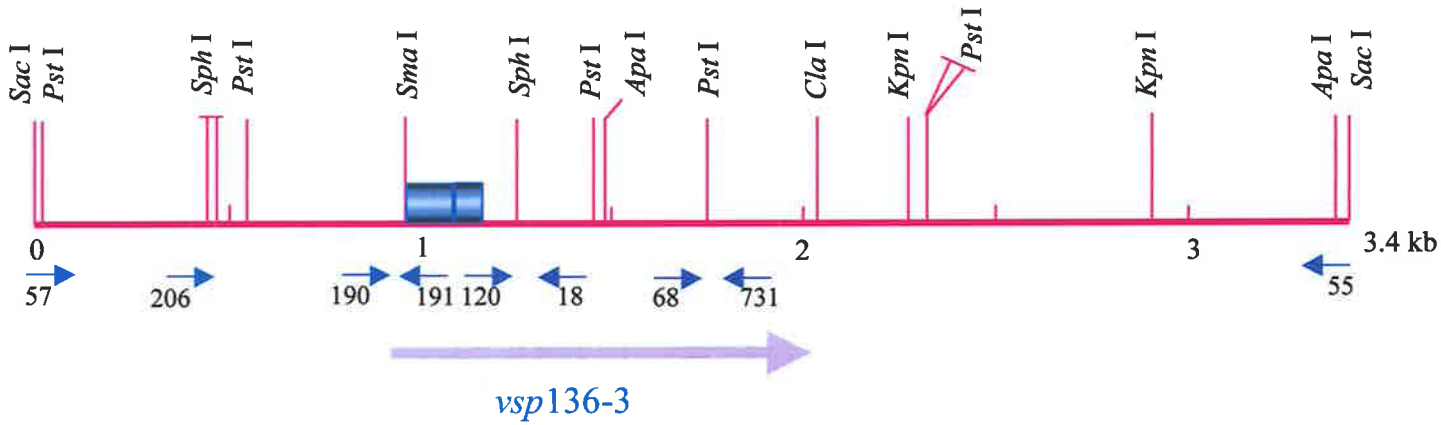


Figure 3.10. Schematic representation of the pM24-1 insert. The open reading frame (*vsp*136-3) is denoted by a thick arrow and the 1.5 tandem repeat units by a purple shaded box. Restriction sites that were identified experimentally by cleavage are shown. The identity and annealing positions of primers used for sequencing are shown by arrowheads.

Figure 3.11. Nucleotide and deduced amino acid sequences specified by the single open reading frame within the characterised 2,768-bp segment of the 3.4-kb insert of pM24-1. The 1,185-bp coding sequence (*vsp*136-3) is indicated. The deduced N-terminal signal peptide and C-terminal invariant segment, 'CRGKA', are underlined, as also is the stop codon, polyadenylation signal sequence (AGTRAA, double underlined) and VSP gene-specific pre-poly(A) signal sequence (CTTAGRT). The sequence has been submitted to GenBank and is available under accession number AF298860.

```
      Sac I
   1  GAGCTCCTCAACGACGAATTGTGGGATGACGGGGGCTGCTGCAGGGGGGCTGGCTGTGTAATCATCGCC
  70  GTCGGCCATGACAGTTTATTCTCAAAAATAAGATCGGGCAATTAGATTGAACGAGTGATTGGGCCGCAC
 139  CCTCAGAGGCTACAAGAGAGCGCAGGCCCCTGCCAGCCTCGCTTCTCGAATCTGTGTTCATGACAGCAG
 208  ACACGTCGTCCCGTACACTGTGGAGCGCCCTTGCATGTCATGGCACCCTGTCTGTCCTGGCCACTCGTG
 277  ACCATCACGCATTACTAGTACTCTGCGCCAGCTCCCTCGACTGTGCAGTCATGGCTATTCCCCATGTGC
 346  ACAGGCCATCTAGCAGCACTCGCCCGCGTTCTCTGCTCCTCCTCCGCCATTCATCCACGCGGGCCATGC
 415  AGGGTCTATTGCCCCTCCAGCGCTGGCCACATTGGTGCTCTCTTCTGAGGCTGCATGCCCTAAAGGATC
 484  GCCTGTGCTCACTGCATTTTTAGAGCCCATCGCACACCGGGGGCTCCCTCGCAATCGGTGCCCAGAAAG
 553  TCACGGGCAGGACTCTTCGGTACGTTTACTACCGAGACACCTTACTCCTGCAGAGGAAGCCGACGAATA
 622  CTCCACAGCAATGTACAGGAGTGCACCCCGAACAATCATTCCACAAGAGAGGTGCGATCTGAGTCTGGC
 691  CCTCACTGGTCCTGTTCCTATCGAAGTCAAGGGGAACTGCACTATCAACTAGCCTCCGATGGGTGTGCT
 760  CCTATCTCGTCCTCCTCCAGTCCAACTGATCAAGAGAGCAAGAGCGGTTGCCTCTCACTGCTCGCGGCA
VSP136-3                                                  M  L  V  G  F  L  L  V
 829  ACACATCAGCCTGATGAAAAAGCAGTAGCACACGAACTCCGGCTAATGTTGGTAGGATTTCTCCTAGTT
   9  C  A  T  V  L  A  K  D  S  G  K  D  T  C  F
 898  TGTGCAACCGTGCTGGCAAAGGACAGCGGGAAGGACACATGCTTC

Repeat region (1 copy):
  24  P  G  Y  T  I  N  T  D  T  K  Q  C  T  K  D  P  E  A  P  C  N  V  E
 943  CCGGGCTACACCATCAACACTGACACGAAGCAGTGCACCAAGGACCCAGAAGCGCCGTGCAACGTCGAG
  47  G  C  E  T  C  V  E  G  N  A  Q  Q  C  K  T  C  R
1012  GGCTGTGAGACCTGCGTGGAGGGCAACGCCCAGCAGTGCAAGACGTGCCGT

+ one half-copy:
  64  P  G  Y  T  I  N  T  D  T  K  Q  C  T  K  D  P  E  A  P  C  N
1063  CCGGGCTACACCATCAACACTGACACGAAGCAGTGCACCAAGGACCCAGAAGCGCCGTGCAAT

Non-repeat 3' segment:
  85                        T  P  N  C  K  T  C  D  N  P  K  T  D  S  E  I
1126                        ACTCCCAACTGTAAGACCTGTGACAACCCCAAAACAGACAGTGAAATC
 101  C  T  E  C  N  D  G  N  Y  L  T  P  T  N  Q  C  V  P  D  C  T  T  I
1174  TGCACTGAATGTAATGACGGCAACTACCTCACCCCTACAAACCAATGCGTACCTGACTGCACGACCATC
 124  S  G  Y  Y  G  D  T  D  K  K  C  K  A  C  N  P  E  C  A  E  C  V  G
1243  AGCGGATACTACGGAGATACTGACAAGAAGTGTAAGGCATGCAACCCTGAGTGCGCTGAGTGCGTTGGG
 147  P  A  N  N  Q  C  S  S  C  P  A  G  K  K  L  T  Y  T  D  D  S  N  P
1312  CCGGCCAACAATCAGTGTAGTTCCTGTCCTGCTGGCAAGAAACTGACATATACAGATGACAGCAATCCT
 170  N  N  G  G  T  C  G  D  A  C  K  V  S  A  D  G  T  G  C  E  T  C  G
1381  AATAACGGAGGCACCTGCGGGGATGCGTGCAAGGTGTCTGCAGATGGCACTGGCTGTGAGACATGCGGG
 193  A  Q  I  G  G  T  A  Y  C  S  K  C  K  T  S  T  Q  A  P  L  N  G  D
1450  GCCCAAATAGGAGGAACTGCATACTGCTCAAAGTGCAAGACCTCCACTCAGGCCCCCTTGAACGGCGAC
 216  C  A  A  S  S  R  A  T  F  C  T  K  M  G  N  G  V  C  T  Q  C  E  D
1519  TGTGCGGCCAGTTCACGAGCAACTTTCTGTACTAAAATGGGTAATGGGGTGTGCACTCAATGTGAAGAT
 239  N  Y  F  L  K  D  G  G  C  Y  Q  T  D  R  Q  P  G  K  Q  V  C  S  S
1588  AACTACTTCCTCAAGGACGGCGGTTGCTATCAAACAGATAGACAGCCTGGCAAGCAAGTGTGTAGTAGT
 262  A  Q  G  G  N  G  K  C  Q  A  C  A  N  G  L  A  A  T  D  G  N  C  A
1657  GCACAGGGAGGCAATGGTAAGTGTCAGGCATGTGCCAATGGCTTAGCAGCAACTGATGGCAACTGTGCA
 285  E  C  H  P  T  C  A  T  C  S  T  A  G  A  A  D  K  C  K  T  C  A  T
1726  GAATGTCATCCTACTTGTGCTACGTGCTCGACTGCAGGTGCGGCTGATAAGTGTAAGACCTGTGCCACT
 308  G  Y  Y  K  E  N  G  D  D  T  T  D  G  P  C  K  K  C  S  E  K  I  S
1795  GGGTATTACAAGGAAAACGGTGACGATACCACTGATGGCCCGTGCAAGAAGTGCTCAGAGAAGATCTCT
```

Figure 3.11 (pM24-1), p. 2

```
 331  G  C  K  Q  C  V  S  S  S  G  S  S  V  I  C  L  E  S  E  V  G  T  G
1864  GGATGCAAGCAATGTGTGTCATCATCTGGCAGCTCAGTCATATGTTTAGAATCAGAAGTAGGCACTGGT
 354  G  S  V  N  K  S  G  L  S  T  G  A  I  A  G  I  A  V  A  V  I  V  V
1933  GGAAGCGTCAACAAGAGCGGCCTCAGCACGGGGGCCATTGCAGGCATCGCTGTGGCTGTCATCGTTGTT

 377  V  G  G  L  V  G  F  L  C  W  W  F  L  C  R  G  K  A  *
2002  GTGGGGGGCCTCGTAGGCTTCCTTTGCTGGTGGTTCCTCTGCCGCGGGAAGGCGTAGGTTTAACTGTGT
2071  ACTTAGGTAGTAAACCGTCATCGATGGGTCTGCTCGGTGTCTGTTCCTGCTAGCACGGACAGAAGGGTT
2140  TCAGCCGGTGCGCTAAGCATCAGGCGTGTGGATGGATGCTCAGTTTATCCAGTAGCACGCCCTGTCCA
2209  AGCTTCACAAGTGACCAACAGTGCTGTACAGGTACCTAGAGACCAGACCGCAGATCCCATGCATTGAAT
2278  GCGGCCCCTCTGCAGCTGCAGGACGGGCCGGTCTGGCAATCTATGATCAAGCAGGAGCCCTGTTCTTGC
2347  AGACCTTGCAGCACTTACAGACCTGCATTGCAGAGAGGCATTCATGCGTCTCCGCAAGGATCGGTGCAT
2416  TGACTGGCGAAGAAAAGAGGGACACTGACAAGGGCTGCCTATACAGCTGCTGATATTGATAACAAGGCA
2485  AGCACACCCAAACACCATGCACTCGCAGGCTGCTACCTTGAAGGGTCGCATGAGCACATATAATGGGGA
2554  CTTGACTCCATCGAAAAAGGGCTGCTTCTCATTTGAACAGAAAACGATTATNGGCCCCACCGCTTGCGT
2623  CATATGCTCCAAAAACAGGCACAGGCTGAACCCNNCANAATCNCNNAACCATCCTCTGGCTGNCTGGGT
2692  TCACGCTTGCCATCCAACAATGAAAGCTTCCCCATNGGCGGNATGGAACCATNCTCTCNTCCCACANCA
2761  AGGGGNAA
```

Table 3.1. Nucleotide sequence similarity of the *vsp* 136-3 locus to other closely-related VSP gene sequences

Gene	*vsp* 136-3			
	5'-UT[a]	5'-NR[b]	Rep[c]	3'-NR[d]
	% Nucleotide identity (length of overlap, bp)[e]			
vsp 136-1	99 [873]	100 [73]	100 [120]	86.6 [960]
vsp 136-2	96 [295]	100 [73]	100 [120]	86.6 [960]
vsp 136-4	99.6 [873]	100 [73]	100 [120]	100 [960]
vsp R2	-	-	78 [116]	76 [644]
vsp 52	85 [177]	95 [73]	78 [98]	87.6 [960]
crp 65	86 [882]	93 [66]	50 [119]	93.8 [960]

[a]. Non coding 5' region

[b]. Non-repeat 5' region

[c]. Repeat region

[d]. Non-repeat 3' region

[e]. Determined by FASTA using DNASIS.

ORF was identified. This represented, like the pM6-1 ORF (section 3.6), an intact vsp gene and on the basis of subsequent comparative analyses, it was designated *vsp*136-3 (**Fig. 3.10**). The VSP encoded by *vsp*136-3 contains 394 amino acid residues, 16 copies of the CXXC motif, a typical N-terminal signal peptide, and C-terminal hydrophobic (membrane-spanning) and invariant (-CRGKA) VSP signature domains (**Fig. 3.11**). An extended polyadenylation signal sequence (CTTAGGTAGTAAA) is situated 14 bp beyond the *vsp*136-3 stop codon. The gene contains only one and a half tandem copies of the *crp*136 120-bp repeat (100% nt sequence identity), in contrast to the 23.5 tandem copies in *crp*136 and *vsp*136-2, but the coding sequence upstream of the repeat elements nevertheless exhibits 100% nt sequence identity with the corresponding (5') segment of *crp*136 (**Fig. 3.16**). The sequence identity between *crp*136 and *vsp*136-3 extends to the upstream *Sac* I site (873 bp upstream from the start codon of *vsp*136-3) which was used to clone this genomic fragment (**Fig. 3.18**). However despite their similarity in these segments, the two genes were found to diverge within the 3' (non-repeat) region of the coding sequence as indicated by the data summarised in **Table 3.1**.

3.7 Characterisation of the pM42-2 insert

The genomic *Sac* I fragments cloned within plasmids pM42-2 and pM49-1 (identified by screening bacterial transformants with the *vsp*R2-R probe; see section 3.4) contained similar sized (≈ 7.5-kb) inserts. The two constructs proved to be identical by restriction analysis, Southern hybridisation and preliminary sequence data. Therefore, pM42-2 was chosen for more detailed characterisation of the insert DNA (e.g. **Fig. 3.12**). A comparison of the restriction profiles of pM6-1 (containing *vsp*136-2, with 23.5 tandem repeats; described in section 3.5) and pM42-2 (which, from comparison with pM24.1 [Fig. 3.9], appeared to contain multiple repeats also) is shown in **Fig. 3.13**. This revealed differences in the restriction profiles of these two cloned fragments. Southern hybridisation analyses of pM42-2

a b c d e f g h i j k l

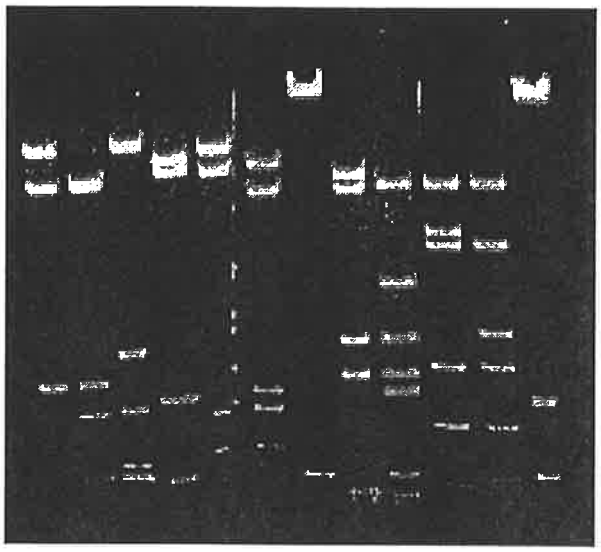


Figure 3.12. Single- and double-enzyme restriction analysis of the pM42-2 construct. Replicate aliquots of plasmid DNA were incubated overnight with different combinations of restriction endonucleases. Fragments were electrophoresed on 1% agarose gel and stained with ethidium bromide. The endonucleases tested were:

a	*Apa* I	g	*Dra* I
b	*Apa* I + *Dra* I	h	*Pst* I
c	*Acc* I + *Dra* I	i	*Pst* I + *Dra* I
d	*Bam* HI + *Dra* I	j	*Sph* I
e	*Sma* I	k	*Sph* I + *Dra* I
f	*Sma* I + *Dra* I	l	*Xho* I + *Dra* I

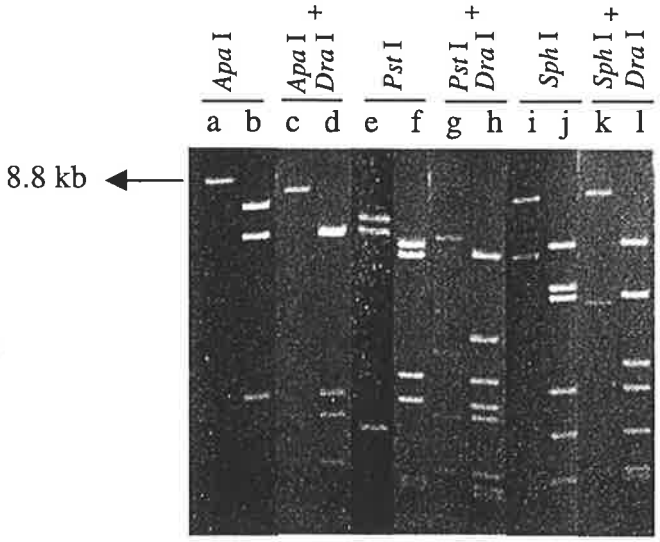


Figure 3.13. Comparison of the restriction fragment profiles of plasmid constructs pM6-1 and pM42-2. Replicate aliquots of plasmid DNA were incubated with different restriction endonucleases. The reaction mixtures were subjected to electrophoresis on 1% agarose and stained with ethidium bromide.

pM6-1: Lanes a, c, e, g, I, k

pM42-2: Lanes b, d, f, h, j, l

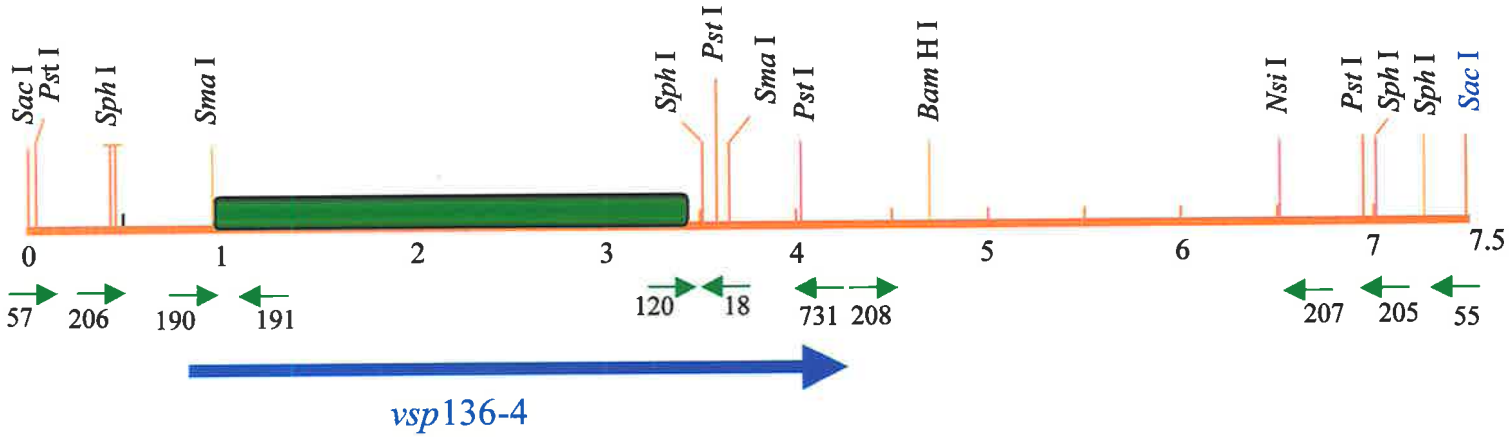


Figure 3.14. Schematic representation of the pM42-2 insert. reading frame (*vsp*136-4) is denoted by a thick arrow and the tandem repeat region by a green shaded box. Restriction sites that were identified experimentally by cleavage are shown. The identity and annealing positions of primers used for sequencing are shown by arrowheads.

Figure 3.15. Nucleotide and deduced amino acid sequences specified by the single open reading frame within the characterised 4,632-bp segment of the insert of pM42-2. The 3,465-bp coding sequence (*vsp*136-4) is indicated. The deduced N-terminal signal peptide and C-terminal invariant segment, 'CRGKA', are underlined, as also is the stop codon, polyadenylation signal sequence (AGTRAA, double underlined) and VSP gene-specific pre-poly(A) signal sequence (CTTAGRT). The sequence has been submitted to GenBank and is available under accession number AF298861.

```
      Sac I
   1  GAGCTCCTCAACGACGAATTGTGGGATGACGGGGGCTGCTGCAGGGGGGCTGGCTGTGTAATCATCGCC
  70  GTCGGCCATGACAGTTTATTCTCAAAAATAAGATCGGGCAATTAGATTGAACGAGTGATTGGGCCGCAC
 139  CCTCAGAGGCTACAAGAGAGCGCAGGCCCCTGCCAGCCTCGCTTCTCGAATCTGTGTTCATGACAGCAG
 208  ACACGTCGTCCCGTACACTGTGGAGCGCCCTTGCATGTCATGGCACCCTGTCTGTCCTGGCCACTCGTG
 277  ACCATCACGCATTACTAGTACTCTGCGCCAGCTCCCTCGACTGTGCAGTCATGGCTATTCCACATGTGC
 346  ACAGGCCATCTAGCAGCACTCGCCCGCGTTCTCTGCTCCTCCTCCGCCATTCATCCACGCGGGGCATGC
 415  AGGGTCTAATGCCCCTCCAGCGCTGGCCACATTGGTGCTCTCTTCTGAGGCTGCATGCCCTAAAGGATC
 484  GCCTGTGCTCACTGCATTTTTAGAGCCCATCGCACACCGGGGGCTCCCTCGCAATCGGTGCCCAGAAAG
 553  TCACGGGCAGGACTCTTCGGTACGTTTACTACCGAGACACCTTACTCCTGCAGAGGAAGCCGACGAATA
 622  CTCCACAGCAATGTACAGGAGTGCACCCCGAACAATCATTCCACAAGAGAGGTGCGATCTGAGTCTGGC
 691  CCTCACTGGTCCTGTTCCTATCGAAGTCAAGGGGAACTGCACTATCAACTAGCCTCCGATGGGTGTGCT
 760  CCTATCTCGTCCTCCTCCAGTCCAACTGATCAAGAAAGCAAGAGCGGTTGCCTCTCACTGCTCGCGGCA
VSP136-4                                                  M  L  V  G  F  L  L  I
 829  ACACATCAGCCTGATGAAAAAGCAGTAGCACACGAACTCCGGCTAATGTTGGTAGGATTTCTCCTAATT
   9  C  A  T  V  L  A  K  D  S  G  K  D  T  C  F
 898  TGTGCAACCGTGCTGGCAAAGGACAGCGGGAAGGACACATGCTTC
```

Repeat region (20 copies):

```
  24  P  G  Y  T  I  N  T  D  T  K  Q  C  T  K  D  P  E  A  P  C  N  V  E
- 943  CCGGGCTACACCATCAACACTGACACGAAGCAGTGCACCAAGGACCCAGAAGCGCCGTGCAACGTCGAG
  47  G  C  E  T  C  V  E  G  N  A  Q  Q  C  K  T  C  R
1012  GGCTGTGAGACCTGCGTGGAGGGCAACGCCCAGCAGTGCAAGACGTGCCGT
```

+ one half-copy:

```
 824  P  G  Y  T  I  N  T  D  T  K  Q  C  T  K  D  P  E  A  P  C  N
3343  CCGGGCTACACCATCAACACTGACACGAAGCAGTGCACCAAGGACCCAGAAGCGCCGTGCAAT
```

Non-repeat 3' segment:

```
 845                        T  P  N  C  K  T  C  D  N  P  K  T  D  S  E  I
3406                        ACTCCCAACTGTAAGACCTGTGACAACCCCAAAACAGACAGTGAAATC
 861  C  T  E  C  N  D  G  N  Y  L  T  P  T  N  Q  C  V  P  D  C  T  T  I
3454  TGCACTGAATGTAATGACGGCAACTACCTCACCCCTACAAACCAATGCGTACCTGACTGCACGACCATC
 884  S  G  Y  Y  G  D  T  D  K  K  C  K  A  C  N  P  E  C  A  E  C  V  G
3523  AGCGGATACTACGGAGATACTGACAAGAAGTGTAAGGCATGCAACCCTGAGTGCGCTGAGTGCGTTGGG
 907  P  A  N  N  Q  C  S  S  C  P  A  G  K  K  L  T  Y  T  D  D  S  N  P
3592  CCGGCCAACAATCAGTGTAGTTCCTGTCCTGCTGGCAAGAAACTGACATATACAGATGACAGCAATCCT
 930  N  N  G  G  T  C  G  D  A  C  K  V  S  A  D  G  T  G  C  E  T  C  G
3661  AATAACGGAGGCACCTGCGGGGATGCGTGCAAGGTGTCTGCAGATGGCACTGGCTGTGAGACATGCGGG
 953  A  Q  I  G  G  T  A  Y  C  S  K  C  K  T  S  T  Q  A  P  L  N  G  D
3730  GCCCAAATAGGAGGAACTGCATACTGCTCAAAGTGCAAGACCTCCACTCAGGCCCCCTTGAACGGCGAC
 976  C  A  A  S  S  R  A  T  F  C  T  K  M  G  N  G  V  C  T  Q  C  E  D
3799  TGTGCGGCCAGTTCACGAGCAACTTTCTGTACTAAAATGGGTAATGGGGTNTGCACTCAATGTGAAGAT
 999  N  Y  F  L  K  D  G  G  C  Y  Q  T  D  R  Q  P  G  K  Q  V  C  S  S
3868  AACTACTTCCTCAAGGACGGCGGTTGCTATCAGACAGATAGACAGCCCGGCAAGCAAGTGTGTAGTAGT
1022  A  Q  G  G  N  G  K  C  Q  T  C  A  N  G  L  A  A  T  D  G  N  C  A
3937  GCACAGGGAGGCAATGGTAAGTGTCAGACATGTGCCAATGGCTTAGCAGCAACTGATGGCAACTGTGCA
1045  E  C  H  P  T  C  A  T  C  S  T  A  G  A  A  D  K  C  K  T  C  A  T
4006  GAATGTCATCCTACTTGTGCTACGTGCTCGACTGCAGGTGCGGCTGATAAGTGTAAGACCTGTGCCACT
1068  G  Y  Y  K  E  N  G  D  D  T  T  D  G  P  C  K  K  C  S  E  K  I  S
4075  GGGTATTACAAGGAAAACGGTGACGATACCACTGATGGCCCGTGCAAGAAGTGCTCAGAGAAGATCTCT
```

Figure 3.15 (pM42-1), p. 2

```
1091  G  C  K  Q  C  V  S  S  S  G  S  S  V  I  C  L  E  S  E  V  G  T  G
4144  GGATGCAAGCAATGTGTGTCATCATCTGGCAGTTCAGTCATATGTTTAGAATCAGAAGTAGGCACTGGT
1114  G  S  V  N  K  S  G  L  S  T  G  A  I  A  G  I  A  V  A  V  I  V  V
4213  GGAAGCGTCAACAAGAGCGGCCTTTCTACTGGCGCCATCGCAGGCATCGCTGTGGCTGTCATCGTTGTT

1137  V  G  G  L  V  G  F  L  C  W  W  F  L  C  R  G  K  A  *
4282  GTGGGGGGCCTCGTAGGCTTCCTTTGCTGGTGGTTCCTGTGCAGAGGAAAAGCATAGATACTTAGGTAG
4351  TAAACGCATCCACTACGTGTGGCCTGTGATACTCAGCTCACCGCGCAGGGAAACAGAAAGTCCACGAGT
4420  AGGCACGGGACCACCCACGTGCTTGCAGCCTCTCTGCTGGCCCTCTTCCGGTGTCTCTACATCCTCAGA
4489  GAGCTGGGGCTTGTGCCTTCCCTGCGCGGATCGACCCCTTGTGGAACTCCGTCTGCCAATGCTCTCTCT
4558  GTCCCTGTGGCATCCTCCTAGACAACANCTGCTGGGCCGGCTACCCGGCAGATACACCGAAAGTTGATC
4627  CTCCCA
```

located the sequence related to the *vsp*R2 repeat within the 5' and central portions of the insert (**Figs. 3.9, 3.14**). It was noticed during the preliminary sequencing from the ends of the insert using T3 and T7 primers that the 5' sequence was nearly identical to the 5' end of the pM24-1 insert. As a result, the sequence of 4,632 bp from the 5' end of the pM42-2 insert was determined (**Fig. 3.15**). In addition, to check for the presence of a second ORF, the sequence of a further 1,700 bp from the 3' end of the insert was also determined (**Fig. 3.14**). This revealed no evidence of a second coding sequence.

The ORF identified in the pM42-2 insert, designated *vsp*136-4, is depicted in **Fig. 3.14** and presented in detail in **Fig. 3.15**. It commences 874 bp from the 5' end of the insert, as does the equivalent ORF in pM24.1, and it encodes a polypeptide consisting of 1,154 amino acids. The similarity between *vsp*136-3 (pM24.1, Fig. 3.11) and *vsp*136-4 (pM42-2) extends from the 5' untranslated region across the first tandem repeat region. However, the two genes differ in their content of tandem repeat units. A comprehensive restriction analysis was performed (as exemplified in **Figs. 3.12 & 3.13**) and the sizes of fragments derived from pM42-2 insert were determined (as in section 3.6) in order to calculate how many tandem copies of the repeat element occur in the *vsp*136-4 gene. These calculations (not shown) indicated that *vsp*136-4 contains 20.5 tandem copies of the 120-bp repeat unit, i.e. three fewer than *crp*136 and *vsp*136-2. The available data are insufficient to determine whether these sequences represent different alleles of the same gene or different loci.

3.8 Definition of a '*vsp*136' gene subfamily

The high level of sequence identity between the three loci (*vsp*136-2, *vsp*136-3 & *vsp*136-4) and a segment of *vsp*R2 identified from the work described in this chapter, together with the level of their similarity to *crp*136, *crp*65 and *vsp*52 (previously characterised vsp genes known to contain tandem repeats), suggested that these loci represent a distinct subset of genes within the vsp gene family. We have designated this subset the '*vsp*136' subfamily because of their similarity to *crp*136, the first reported member of this

group of genes, which was expressed by a subclone of the WB isolate (Chen *et al.* 1995). As the prototype locus for the *vsp*136 gene subfamily, we have tentatively redesignated the *crp*136 locus as *vsp*136-1. In describing the *crp*136 and *crp*65 genes (Chen *et al.* 1996), Upcroft *et al.* (1997) proposed that the two loci belong to a gene subfamily ('*crp*136') whose members possess variable-length segments consisting of different tandem repeat elements within a highly conserved 'cassette'. Surprisingly, they failed to mention that *crp*136 and *crp*65 encode polypeptides that are typical VSP. Although the identification of CRP136 and CRP65 as VSP is unambiguous (determined primarily by the characteristic, highly conserved C-terminal segment and secondly by the 12-13mol% cysteine content and multiple '-CXXC-' motifs), Chen *et al.* (1995, 1996) and Upcroft *et al.* (1997) instead proposed that the encoded proteins are toxins.

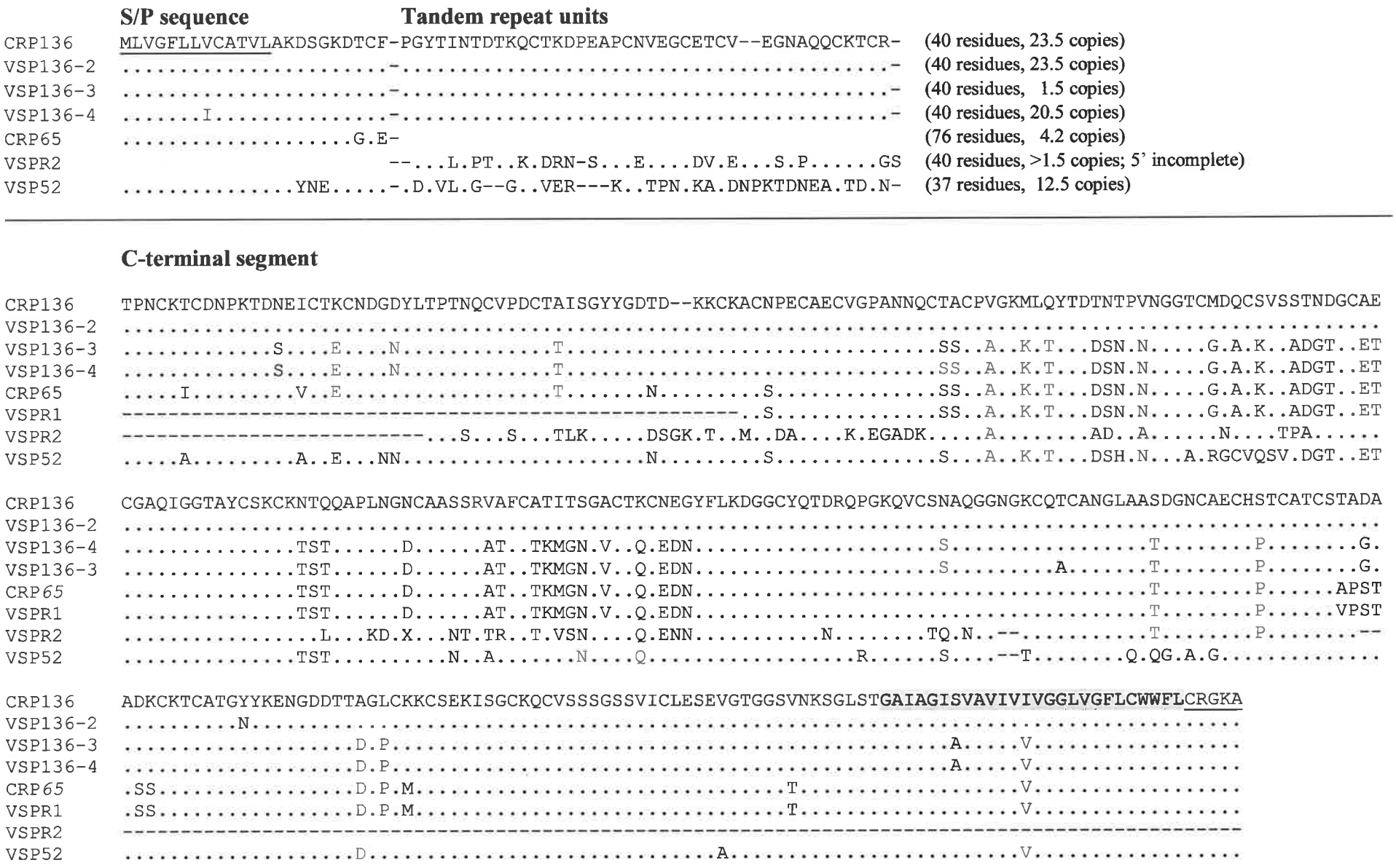
In the present study, three related genes have been characterised within the genome of a distinct isolate, Ad-1/c3, that belongs to the same genetic subtype of *G. intestinalis* (type A-I) as the WB isolate. These newly-defined genes encode the same tandem repeat sequence as that reported for *vsp*136-1 (*crp*136), with *vsp136*-2 possessing the same number of tandem copies (23.5) as *vsp*136-1 whilst *vsp*136-3 and *vsp*136-4 contain fewer copies (1.5 and 20.5 respectively). They were found to contain 5' non-repeat and repeat coding regions that were similar to *vsp*136-1 (*crp*136), but their 3' non-repeat coding segments are different from *vsp*136-1.

3.9 The VSP encoded by *crp*136-like (*vsp*136 subfamily) genes

An alignment of the amino acid sequences encoded by these different *crp*136-like loci (**Fig. 3.16**) highlights the overall close similarity of the polypeptides. This similarity is evident in the N- and C-terminal segments, including the N-terminal non-repeat and central tandem repeat regions. Four of the inferred proteins, VSP136-1 (CRP136), VSP136-2, VSP136-3 and VSP136-4 contain the same (identical) tandem repeat element but possess

Figure. 3. 16. Aligned inferred amino acid sequences encoded by the *vsp*136 loci. Gaps (-) introduced to maximise the alignments. Dotes indicate amino acid identity between the CRP136 (top row) and corresponding positions in other proteins. The presumptive membrane-spanning hydrophobic segment is shaded in grey. The predicted N-terminal signal peptide and invariant hydrophobic segment at the C-terminus, underlined. Only one repeat sequence is shown for each protein. The CRP65 repeat is not shown due to much longer sequence and lack of any significant similarity with other aligned repeat sequences. Motif identities shared between more than two polypeptides (which are different from CRP136), are highlighted in red.

Figure 3.16


S/P sequence
Tandem repeat units
CRP136 MLVGFLLVCATVLAKDSGKDTCF-PGYTINTDTKQCTKDPEAPCNVEGCETCV--EGNAQQCKTCR- (40 residues, 23.5 copies)
VSP136-2 .....-..... (40 residues, 23.5 copies)
VSP136-3 .....-..... (40 residues, 1.5 copies)
VSP136-4 .....I.....-..... (40 residues, 20.5 copies)
CRP65 .....G.E- (76 residues, 4.2 copies)
VSPR2 --...L.PT..K.DRN-S...E....DV.E...S.P.....GS (40 residues, >1.5 copies; 5' incomplete)
VSP52 .....YNE.....-.D.VL.G--G..VER---K..TPN.KA.DNPKTDNEA.TD.N- (37 residues, 12.5 copies)
C-terminal segment
CRP136 TPNCKTCDNPKTDNEICTKCNDGDYLTPTNQCVPDCTAISGYYGDTD--KKCKACNPECAECVGPANNQCTACPVGKMLQYTDTNTPVNGGTCMDQCSVSSTNDGCAE
VSP136-2 .....
VSP136-3 .....S....E....N.....T.....SS..A..K.T...DSN.N.....G.A.K..ADGT..ET
VSP136-4 .....S....E....N.....T.....SS..A..K.T...DSN.N.....G.A.K..ADGT..ET
CRP65 .....I.....V..E.....T.....N.....S.....SS..A..K.T...DSN.N.....G.A.K..ADGT..ET
VSPR1 -----..S.....SS..A..K.T...DSN.N.....G.A.K..ADGT..ET
VSPR2 -----...S...S...TLK.....DSGK.T..M..DA....K.EGADK.....A.....AD..A.....N....TPA.....
VSP52 .....A.....A..E...NN.....N.....S.....S...A..K.T...DSH.N...A.RGCVQSV.DGT..ET
CRP136 CGAQIGGTAYCSKCKNTQQAPLNGNCAASSRVAFCATITSGACTKCNEGYFLKDGGCYQTDRQPGKQVCSNAQGGNGKCQTCANGLAASDGNCAECHSTCATCSTADA
VSP136-2 .....
VSP136-4 .....TST.....D.....AT..TKMGN.V..Q.EDN.....S.....T.....P.....G.
VSP136-3 .....TST.....D.....AT..TKMGN.V..Q.EDN.....S.....A.....T.....P.....G.
CRP65 .....TST.....D.....AT..TKMGN.V..Q.EDN.....T.....P.....APST
VSPR1 .....TST.....D.....AT..TKMGN.V..Q.EDN.....T.....P.....VPST
VSPR2 .....L...KD.X...NT.TR..T.VSN....Q.ENN.....N.....TQ.N..--.....T.....P.....--
VSP52 .....TST.....N..A.....N....Q.....R.....S....--T.....Q.QG.A.G.....
CRP136 ADKCKTCATGYYKENGDDTTAGLCKKCSEKISGCKQCVSSSGSSVICLESEVGTGGSVNKSGLSTGAIAGISVAVIVIVGGLVGFLCWWFLCRGKA
VSP136-2 .....N.....
VSP136-3 .....D.P.....A.....V.....
VSP136-4 .....D.P.....A.....V.....
CRP65 .SS.....D.P.M.....T.....V.....
VSPR1 .SS.....D.P.M.....T.....V.....
VSPR2 -----
VSP52 .....D.....A.....V.....

different numbers of the repeat unit within their tandem arrays, as indicated in **Fig. 3.16**. VSP136-2 is almost identical to VSP136-1 (CRP136), possessing the same number of tandem repeats (23.5 copies) and exhibiting amino acid identity with VSP136-1 at 99.9% of positions along its entire length. There is a very high level of similarity between the non-repeat C-terminal segment of all of the aligned proteins as shown in **Fig. 3.16**. Interestingly, many of the substituted residues (red lettering; **Fig. 3.16**) are common to VSP136-3, VSP136-4, CRP65, VSPR1, VSP52 and to a lesser extent, VSPR2, i.e. all except VSP136-1 and VSP136-2.

There are several possibilities to explain the formation of these two related but nonetheless distinct protein subsets. Ancestral gene duplications followed by subsequent additional duplications of some copies and intergenic recombinations between different segments of the duplicated genes, together with the accumulation of single 'point' mutations, could be expected to result in the appearance of a set of paralogous genes such as these. As is exemplified for *vsp*136-3 in **Table 3.1**, the entire 5' non-repeat and repeat coding region exhibits 99-100% nt sequence identity with the corresponding portion of *vsp*136-1 (*crp*136), *vsp*136-2 and *vsp*136-4. However, in the 3' non-repeat coding segment, *vsp*136-3 is more similar to *vsp*136-4 and *crp*65 (identical at 100% and 93.8% of nt sites respectively). This appears to represent evidence of homologous recombination in different segments of these replicated loci, resulting in the *vsp*136-3 locus having a 5' segment that is nearly identical with one subset of *vsp*136-like genes but a 3' segment that exhibits a similar level of near-identity with a different subset of related *vsp*136 loci.

3.10 Phylogenetic analysis of *vsp*136 loci

To further investigate the differences between the *vsp*136 loci, their phylogenetic relationships were investigated. The multiple alignment of the C-terminal amino acid sequences alignment of the deduced VSP136 polypeptides, generated using CLUSTAL W

(Thompson *et al.* 1994) and shown in Fig. 3.16, was used to generate a γ-2 distance matrix and subjected to Neighbour-Joining method analysis (as implemented in MEGA, Version 1.01; Kumar *et al.* 1994). The outcome of these computations is depicted in **Fig. 3.17**, from which four distinct lineages are evident. The most distinctive is that defined by VSPR2. A second lineage is defined by VSP136-1 and VSP136-2. This forms a cluster that is well resolved from the third lineage, defined by four sequences which form two closely related clusters (VSP136-3 and VSP136-4; CRP65 and VSPR1. The fourth lineage is defined by VSP52. In view of the near-identity of VSP136-1 (identified in the WB isolate) and VSP136-2 (identified in the Ad-1 isolate, belonging to the same genetic subtype as WB), it seems likely that these sequences (VSP136-1 and VSP136-2) represent the same locus. Similarly, VSPR1 (from Ad-1) and CRP65 (from WB) may represent a second, single locus. In contrast, all of the other sequences were identified within the Ad-1/c3 genome or in cDNA derived from Ad-1/c3 mRNA, making it likely that each represents a distinct locus. In all, the data reveal what seem likely to be six paralogues: *vsp*136-1/*vsp*136-2, *vsp*R1/*crp*65, *vsp*R2, *vsp*52, *vsp*136-3 and *vsp*136-4.

3.11 Comparison of 5' and 3' noncoding (flanking) sequences

3.11.1 The 5' non-coding region

Potential promoter sequences have been identified in some 'housekeeping' genes in *Giardia* (e.g. Yee *et al.* 2000), but as yet there has been no successful identification of regulatory element associated with vsp genes. Detection of conserved sequence motifs in the flanking regions of vsp genes may help to identify promoters if they exist, as well as other sequence motifs which might act to facilitate recombinations, e.g. for gene translocations.

To examine the flanking sequences of the *vsp*136 genes for such motifs, an alignment of the 5' non-coding segments of *vsp*136-2, *vsp*136-3, *vsp*136-4, together with the

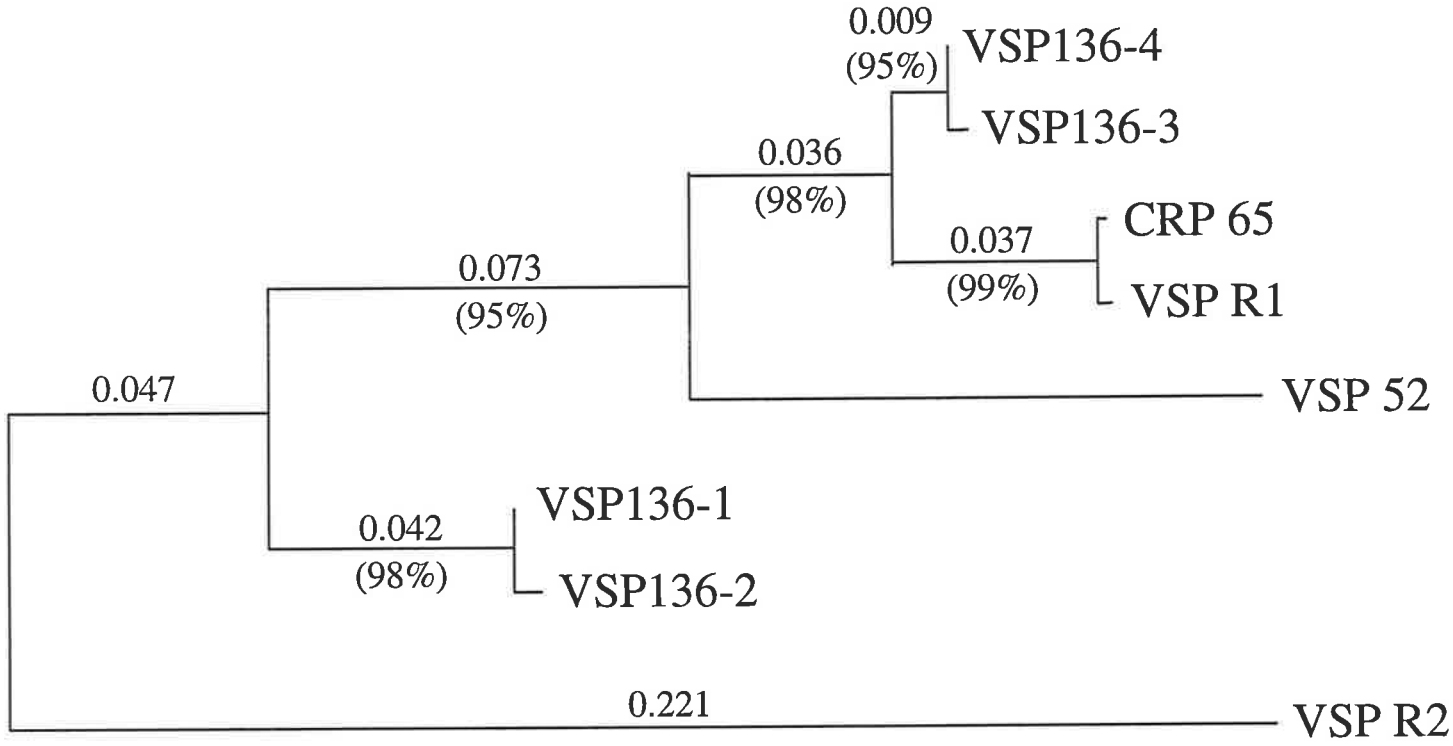
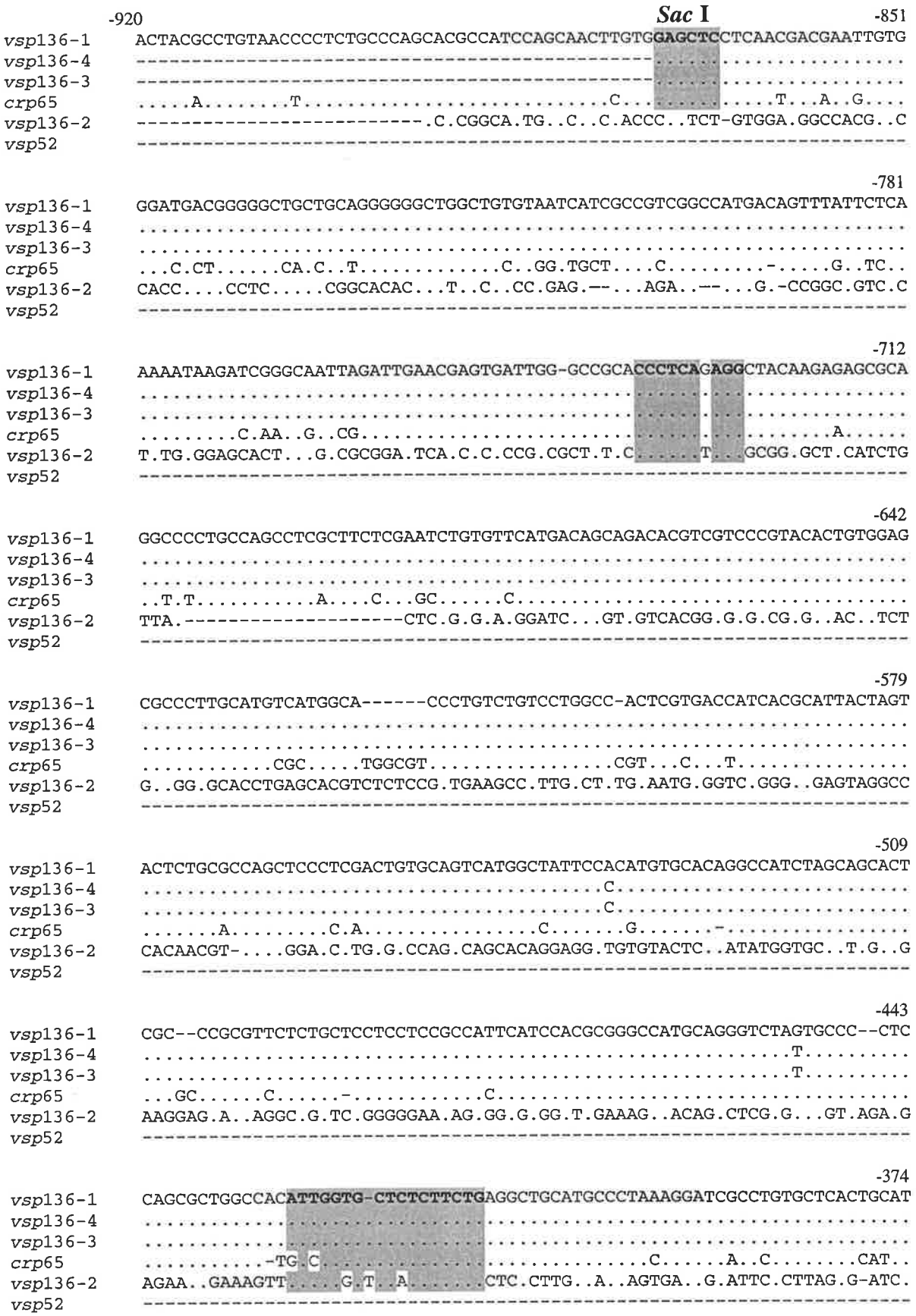


Figure 3-17. Relationships between VSP136-like polypeptides inferred by phylogenetic (Neighbour-Joining) analysis of deduced C-terminal amino acid sequences. Analysis of the repeat-free, 314-residue C-terminal Segments (each commencing beyond the last tandem repeat unit; Fig. 3-16, lower panel), was undertaken using MEGA (Kumar et al. 1994) using the γ-2 distance measure and pairwise deletions. Branch lengths (γ-2 distance) and confidence values (%) for each cluster, determined by bootstrap analysis from 5,000 iterations, are indicated.

corresponding flanking sequences of the three previously characterised genes *vsp*136-1 (*crp*136) and *crp*65 (Chen *et al.* 1995, 1996) and *vsp*52 (Ey *et al.*, unpubl. data), was generated using CLUSTAL W (Thompson *et al.* 1994). This initial alignment required manual editing, the result of which is shown in **Fig. 3.18**. An equivalent alignment of the 3' non-coding segments is presented in **Fig. 3.19**.

In the case of the 5' flanking regions (**Fig. 3.18**), it can be seen that the near-identity of *vsp*136-3 and *vsp*136-4 with *vsp*136-1 continued at least 873 bp upstream from the respective start codons to the common *Sac* I restriction site which, for *vsp*136-3 and *vsp*136-4, constituted the 5' ends of genomic fragments in which these genes were cloned. Across this segment, only four nucleotide substitutions were evident between these three loci – at aligned positions –251 (T/C), -260 (G/C), -451 (T/G) and –536 (C/A). In the case of the *vsp*136-2 gene, which has a coding sequence identical to that of *vsp*136-1 (*crp*136), the sequence of the first 270 bp upstream from the start codon is very similar to that of *vsp*136-1. However, beyond this point and especially beyond nt -300, the sequences diverge significantly (**Fig. 3.18**) – indicating that one of these two genes may have undergone a recombination near this position (nt –300). The 5' flanking sequence of *crp*65 showed occasional nucleotide differences from *vsp*136-1 but overall the two are very similar even 875 bp from the initiation codon (**Fig. 3.18**). Only 177 bp of sequence was available for the 5'flanking region of the *vsp*52 gene. However, this also is very similar to the corresponding segment of *vsp136*-1 (85% identity, or 27 differences over 177 bp).

With respect to more conserved sequence motifs within this 5' non-coding region, it is interesting to note the existence of two highly conserved sequence elements (nts –727 to –736 and –413 to –429; shaded grey, **Fig. 3.18**). These are apparent predominantly because they are evident in the most divergent sequence, *vsp*136-2. These two elements are situated upstream of the more conserved 310-bp segment that lies immediately upstream from the


Sac I
-920
-851
vsp136-1 ACTACGCCTGTAACCCCTCTGCCCAGCACGCCATCCAGCAACTTGTGGAGCTCCTCAACGACGAATTGTG
vsp136-4
vsp136-3
crp65 .....A.....T.....C.....T...A..G....
vsp136-2 ----------.C.CGGCA.TG..C..C.ACCC..TCT-GTGGA.GGCCACG..C
vsp52
-781
vsp136-1 GGATGACGGGGGCTGCTGCAGGGGGGCTGGCTGTGTAATCATCGCCGTCGGCCATGACAGTTTATTCTCA
vsp136-4
vsp136-3
crp65 ...C.CT.....CA.C..T.....C..GG.TGCT....C.....-.....G..TC..
vsp136-2 CACC....CCTC.....CGGCACAC...T..C..CC.GAG.--...AGA..--...G.-CCGGC.GTC.C
vsp52
-712
vsp136-1 AAAATAAGATCGGGCAATTAGATTGAACGAGTGATTGG-GCCGCACCCTCAGAGGCTACAAGAGAGCGCA
vsp136-4
vsp136-3
crp65 .....C.AA..G..CG.....A.....
vsp136-2 T.TG.GGAGCACT...G.CGCGGA.TCA.C.C.CCG.CGCT.T.C.....T...GCGG.GCT.CATCTG
vsp52
-642
vsp136-1 GGCCCCTGCCAGCCTCGCTTCTCGAATCTGTGTTCATGACAGCAGACACGTCGTCCCGTACACTGTGGAG
vsp136-4
vsp136-3
crp65 ..T.T.....A....C...GC.....C.....
vsp136-2 TTA.----------CTC.G.G.A.GGATC...GT.GTCACGG.G.G.CG.G..AC..TCT
vsp52
-579
vsp136-1 CGCCCTTGCATGTCATGGCA------CCCTGTCTGTCCTGGCC-ACTCGTGACCATCACGCATTACTAGT
vsp136-4
vsp136-3
crp65 .....CGC.....TGGCGT.....CGT...C...T.....
vsp136-2 G..GG.GCACCTGAGCACGTCTCTCCG.TGAAGCC.TTG.CT.TG.AATG.GGTC.GGG..GAGTAGGCC
vsp52
-509
vsp136-1 ACTCTGCGCCAGCTCCCTCGACTGTGCAGTCATGGCTATTCCACATGTGCACAGGCCATCTAGCAGCACT
vsp136-4 .....C.....
vsp136-3 .....C.....
crp65 .....A.....C.A.....C.....G.....-.....
vsp136-2 CACAACGT-....GGA.C.TG.G.CCAG.CAGCACAGGAGG.TGTGTACTC..ATATGGTGC..T.G..G
vsp52
-443
vsp136-1 CGC--CCGCGTTCTCTGCTCCTCCTCCGCCATTCATCCACGCGGGCCATGCAGGGTCTAGTGCCC--CTC
vsp136-4 .....T.....
vsp136-3 .....T.....
crp65 ...GC.....C.....-.....C.....
vsp136-2 AAGGAG.A..AGGC.G.TC.GGGGGAA.AG.GG.G.GG.T.GAAAG..ACAG.CTCG.G...GT.AGA.G
vsp52
-374
vsp136-1 CAGCGCTGGCCACATTGGTG-CTCTCTTCTGAGGCTGCATGCCCTAAAGGATCGCCTGTGCTCACTGCAT
vsp136-4
vsp136-3
crp65 .....-TG.C.....C.....A..C.....CAT..
vsp136-2 AGAA..GAAAGTT.....G.T..A.....CTC.CTTG..A..AGTGA..G.ATTC.CTTAG.G-ATC.
vsp52
-305

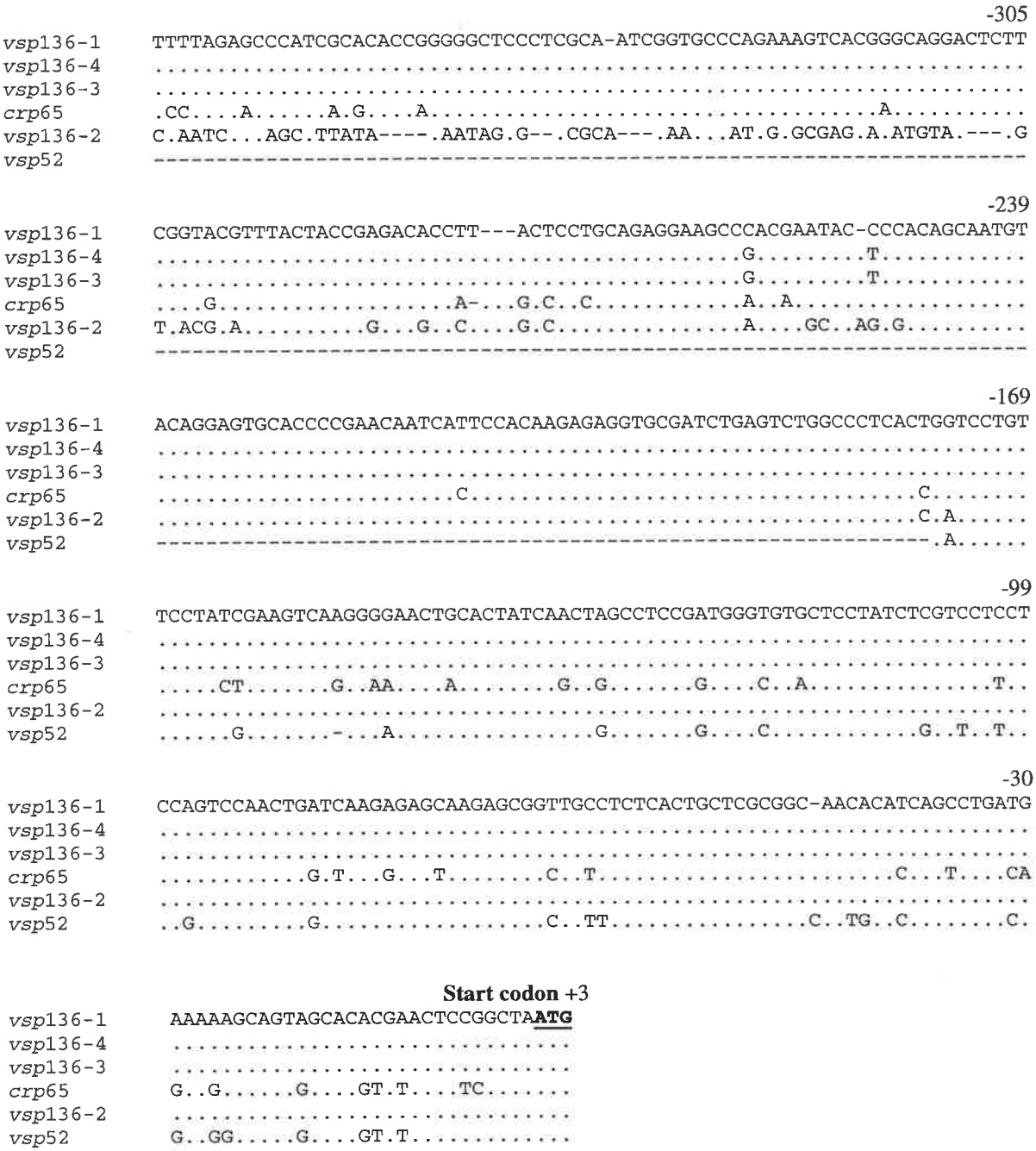


Figure 3. 18. Alignment of the nucleotide sequences from the 5' untranslated regions of all known *crp*136-like ('*vsp*136' subfamily) genes.

Dots represent nucleotide identity with the corresponding positions in v*sp*136-1 (*crp*136, row 1). The start codon is highlighted. Conserved distal nucleotide sequences (between *vsp*136-1, *vsp*136-2, *vsp*136-3, *vsp*136-4 and *crp*65) are shaded in grey, as also is the common *Sac* I cloning site of the genomic fragments containing *vsp*136-1, *vsp*136-3, *vsp*136-4 and *crp*65.

start codon. Their conservation, relative to the surrounding sequence, could indicate a functional role in regulating the expression of these genes. Alternatively, they may function as sites for recombination or the translocation of these genes to an expression site.

3.11.2 The 3' noncoding region

An alignment of the 3' flanking sequences of six *vsp*136 subfamily loci is shown in **Fig. 3.19**. A conserved segment is evident, extending from the stop codon through to the polyadenylation signal motif. Beyond the latter signal motif, only *vsp*136-2 shows significant sequence similarity with *vsp*136-1 (*crp*136). This near-identity (only two nt substitutions in a 143-bp overlap) continues to the *Sac* I cloning site, which is polymorphic (present in *vsp*136-2 but not in *vsp*136-1).

Among the other loci represented in the alignment, beyond the polyadenylation signal motif only *vsp*136-4 and *vsp*52 exhibit a significant degree of sequence identity (**Fig. 3.19**). Surprisingly, the 3' noncoding segments of these two loci are almost identical over the entire 200 bp for which sequence data are available, indicative of either a common recent ancestry or perhaps gene conversion. This similarity between *vsp*136-4 and *vsp*52 continues across the non-repeat 3' portion of their coding sequences, but it does not extend into the tandem repeat region (c.f. **Fig. 3.16**). Interestingly, across the coding sequence *vsp*136-4 shows greater similarity with *vsp*136-3 and *crp*65 than with *vsp*52 (c.f. **Fig. 3.16**) but beyond the polyadenylation signal neither *vsp*136-3 nor *crp*65 shows much sequence similarity with *vsp*136-4 (**Fig. 3.19**). The 3' non-coding region of *crp*65 differs from both pairs (*vsp*136-1/*vsp*136-2 and *vsp*136-4/*vsp*52), but there is residual evidence at some nucleotide sites of low level similarity with *vsp*136-1. The 3'-noncoding region of *vsp*136-3 shows little similarity with any of the other 5 loci beyond the polyadenylation signal motif (**Fig. 3.19**).

Figure 3.19. Nucleotide sequence alignment of the 3' untranslated regions of *crp*136-like loci. Dots represent nucleotide identity with the corresponding positions of *vsp*136-1 (*crp*136, row 1). The stop codon is highlighted, as also is the extended polyadenylation signal sequence (underlined), and the *Sac* I site (only present in *vsp*136-2 shaded in red). Identities between *vsp*136-4 and *vsp*52 are highlighted in red.

3.12 Detection of vsp gene transcripts in *Giardia* trophozoites

The mechanism responsible for VSP switching in *Giardia* has not been elucidated. Neither the frequency of expression of particular vsp genes (relative to other vsp genes), nor whether vsp genes that occur as multiple copies within the genome are expressed more frequently compared with single-copy vsp genes, is known. Having discovered in this project a copy (*vsp*136-2) of the *vsp*136-1 (*crp*136) gene in the Ad-1/c3 *G. intestinalis* genome, as well as two closely-related additional loci (*vsp*136-3, *vsp*136-4), it was of interest to determine whether these '*vsp*136' subfamily genes are expressed more frequently in axenic cultures than known single-copy vsp genes. As mentioned earlier (section 3.10), because *vsp*136-1 and *vsp*136-2 have been identified in different isolates of *G. intestinalis* (but of the same subtype) they may in fact represent the same locus.

In situ mRNA hybridisation was used to quantify the number of cells that expressed *vsp*136 subfamily loci in a long-term (aged) culture of the Ad-1/c3 isolate, i.e. one in which antigenic variants are most likely to predominate and exist in equilibrium with each other. It is of course possible that axenic culture selects for variants expressing particular vsp genes over cells expressing other, less-favourably selected vsp genes. I know of no published reports of the use of *in situ* mRNA hybridisation to detect the expression of vsp genes (or any other type of protein-encoding gene) in *Giardia* trophozoites. However, rDNA (and rRNA) sequences have been detected within the nuclei (or cytoplasm) by *in situ* hybridisation by Kabnick & Peattie (1990) and Macechko *et al.* (1998).

The unique advantage of *in situ* mRNA hybridisation is that individual cells containing transcripts from a given gene can be identified and enumerated within a large population of cells, using single-stranded anti-sense probes. The use of the complementary ('sense') probes which, like the anti-sense probes, should hybridise with the target gene (and possibly related genes) in the nucleus but (unlike anti-sense probes) not with RNA

transcripts, provides an indicator of strand specificity (detection of mRNA). This technique has been in use in this laboratory for the past 18 months and it has given outstanding results.

The hypothesis upon which these experiments rested was that in a *Giardia* culture that has been grown for a long period of time, all possible antigenic variants will exist, presumably in equilibrium with each other and in similar frequencies providing no particular variants are advantaged by the culture medium and conditions. In this idealised situation, the chance and frequency of expression for each particular VSP should be equal. On this basis, by counting the number of cells expressing a particular vsp gene within the cell population and comparing this frequency with that of cells identified as expressing a known single-locus vsp gene, an estimate of the 'copy number' (single, or multiple) of the test genes within the genome can be derived.

For the current study, three different segments of the *vsp*136-2 gene were obtained by PCR amplification. These represented:

1. The repeat region (including the short non-repeat 5' segment), amplified using oligos 190 + 191.
2. The repeat region only, amplified using oligos 120 + 191 (c.f. **Fig. 3.2**).
3. The non-repeat 3' coding region, amplified using oligos 129 + 121 (c.f. **Fig. 3.2**).

After purifying these three amplified products on agarose gels, they were used as templates in single-primer PCR to produce single-stranded DIG-labelled probes. Oligos 121 and 191 were used (separately) to make antisense probes from all three aforementioned products and oligos 190 and 120 were used to make 'sense' probes from the three templates. All of these probes were used subsequently for *in situ* mRNA hybridisation. *Vsp*417-6 is known from Southern hybridisations (Ey & Darby, unpubl. data) to be a single locus in type A-I (group 1) isolates of *G. intestinalis* and it was used as an indication for the number of cells expressing a single-locus vsp gene in a long-term culture. A 300-bp product was amplified from the central region of the cloned *vsp*417-6 gene using primers 25 and 26

(Table 2.1) and this was used to make DIG-labelled single-stranded sense (using oligo 26) and antisense (using oligo 25) probes. Trophozoites collected from confluent tubes of a long-term (>6 months) culture of the Ad-1 isolate were used to test for the presence (and number) of cells staining with probes derived from *vsp*417-6 and different segments of *vsp*136-2. The *in situ* mRNA hybridisation method is described in section 2.14. After hybridisation and overnight staining, the trophozoites were examined by light microscopy. In order to obtain accurate estimates of the percentage of cells that contained mRNA that hybridised with each probe (i.e. those expressing a particular gene), the number of cells stained within a minimum count of 10,000 cells was determined.

As expected, no trophozoites were stained using any of the 'sense' probes in negative control slides, thus verifying the specificity of the assay for gene transcripts (**Fig. 3.20 a**). Examples of cells stained on slides incubated with anti-sense probes are shown in **Fig. 3.20 b, c**. Using these latter probes, approximately 0.57% of the trophozoites contained transcripts that hybridised with the short non-repeat 5' segment of *vsp*136-1 or *vsp*136-2 (derived from the primer 190-191 template; this included a portion of the first repeat unit), whilst 0.68% contained transcripts that hybridised with the tandem repeat probe. The frequency of these variants far exceeded the 0.04% of cells that contained transcripts which hybridised with the non-repeat 3' region probe (**Fig. 3. 20 b, c**). Surprisingly, the [*vsp*417-6]-specific anti-sense probe (known to be a single locus in type A-I *G. intestinalis*) stained more trophozoites (0.8%). This may indicate that *in vitro* conditions favour cells producing some particular VSP and that even in such long-term grown cultures, the expression rate for each vsp gene may not be equal.

At face value, the similar incidence of variants expressing *vsp*417-6 (0.8%), *vsp*136 subfamily genes containing the *crp*136 repeat-(0.68%) and *vsp*136 subfamily genes with related 5' segments (0.57%) suggests that only a single member of the *vsp*136 gene subfamily is functional. However, with no evidence to indicate that any of the newly

Figure 3.20. Detection of *vsp136* subfamily gene expression in variant Ad-1/c3 *Giardia intestinalis* trophozoites by *in situ* mRNA hybridisation. DIG-labelled, single-stranded 'antisense' probes, specific for sequences within the short nonrepeat 5' and/or repeat portion of *crp136*, *vsp136-2,* vsp136-*3 and* vsp136-*4* (Fig. 3.10), were used to identify cells that contained related transcripts (see M&M, section 2.14). Specificity controls included samples exposed to the complementary 'sense' probes. The examples shown include:

Plate (a). ***Negative control, using a 'sense' probe with specificity for the 5 non-repeat and repeat element of the crp136 gene.*** The single-stranded probe was synthesised using primer 190 from a double-stranded template that had been produced in PCR using primers 190 and 191 (c.f. Fig. 3.10). It should not have hybridised with mRNA. No stained cells were detected.

a

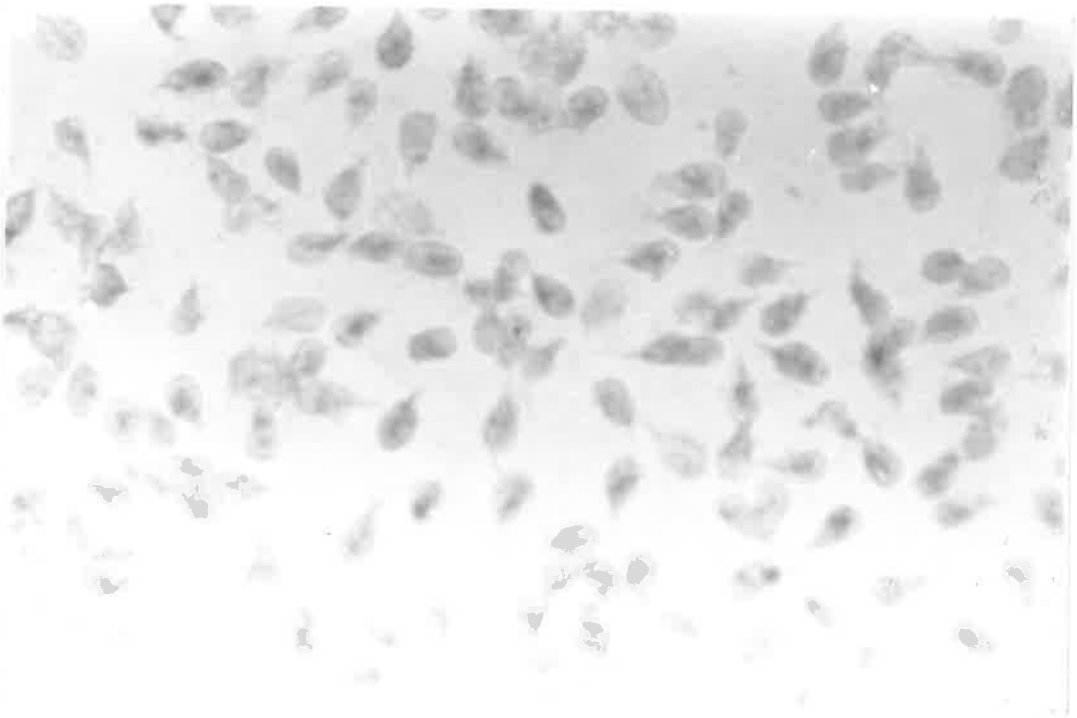
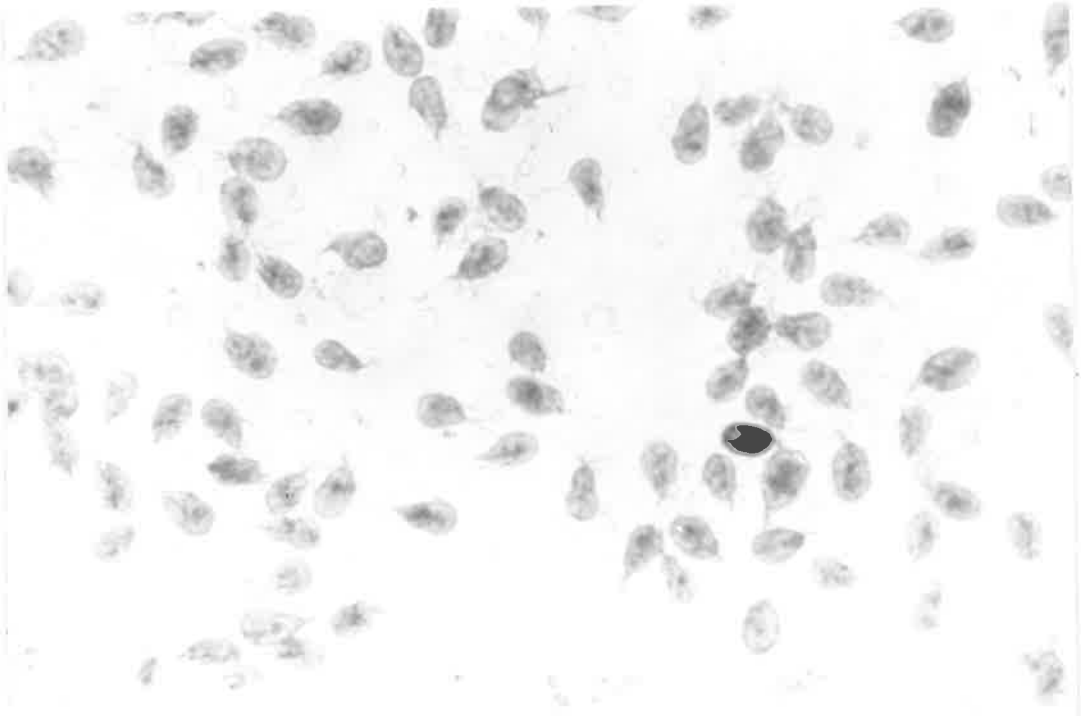


Figure 3.20. Detection of *vsp136* subfamily gene expression in variant Ad-1/c3 *Giardia intestinalis* trophozoites by *in situ* mRNA hybridisation. DIG-labelled, single-stranded 'antisense' probes, specific for sequences within the short nonrepeat 5' and/or repeat portion of *crp136*, *vsp136-2,* vsp136-*3 and* vsp136-*4* (Fig. 3.10), were used to identify cells that contained related transcripts (see M&M, section 2.14). Specificity controls included samples exposed to the complementary 'sense' probes. The examples shown include:

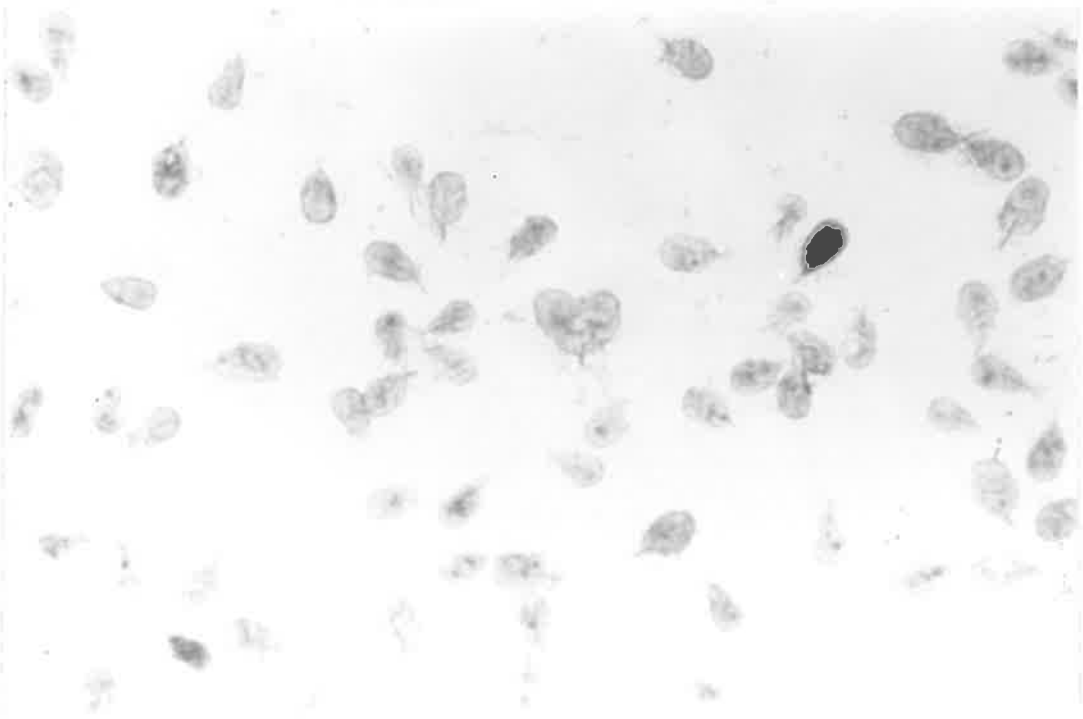
Plate (b). *'Anti-sense' probe specific for 'vsp136'-like genes.* This probe was the complement of the 'sense' probe used for plate (a). It was synthesised from the same template DNA (primers 190 → 191; c.f. Fig. 3.10). Heavy cytoplasmic staining was observed in a minority of cells.

Plate (c). *'Anti-sense' probe specific for 'vsp136'-like genes.* This probe was synthesised using primer 191 from a double-stranded template that had been produced in PCR using primers 120 and 191. Heavy cytoplasmic staining was observed in a minority of cells.

b



c



characterised members of this gene subfamily are defective, this conclusion clearly depends on whether the variants identified in these experiments were indeed all present at an unbiased, equilibrium frequency. At present, this is unknown.

3.13 Discussion

The number of vsp genes that have been completely characterised in the published literature is limited and the majority lack tandem repeat elements. Only four completely characterised vsp genes containing tandem repeat elements (*vsp*A6, *vsp*C5, *crp*136 and *crp*65) had been described prior to this current work (Adam *et al.* 1988b, 1992; Chen *et al.* 1995, 1996) and many issues concerning the stability, number of copies, expression and frequency of recombination of these genes remain poorly understood. These genes (*vsp*A6, *vsp*C5, *crp*136 and *crp*65) were discovered in various subclones of the WB isolate, which belongs to the A-I (group 1) subtype of *G. intestinalis* (Nash *et al.* 1985; Nash & Mowatt 1992a; Ey *et al.* 1992, 1993a,b). However, only a limited amount of work has been done to examine the stability of these tandem repeat-containing genes in other *Giardia* isolates, even among those of the same subtype (Mowatt *et al.* 1994; Yang *et al.* 1994; Yang & Adam 1994, 1995). It was therefore uncertain if these loci were stable and exist unchanged in other *Giardia* isolates of the same genotype or whether they were relatively unstable, perhaps prone to recombinations with related genes or sequences elsewhere in the genome.

Allelic copies of *vsp*A6 (*vsp*A6.2, *vsp*A6.3) and *vsp*C5 (*vsp*C5-S1, *vsp*C5-S2) have been described (Yang & Adam 1994, 1995; Yang *et al.* 1994) in the genome of the WB isolate. However, the complete coding sequences of these copies were not reported and despite the availability of some long-range mapping data, there is no published detailed information about the 3' segments of these vsp genes and whether these differ from the corresponding segment of *vsp*A6. Two other fully characterised vsp genes, *crp*136 and *crp*65, had been found to possess similar 5' and 3' non-repeat regions but unrelated tandem repeats

(Chen *et al.* 1995, 1996). However, no information was available on the full coding and flanking sequences of vsp genes that possess identical repeat elements.

In this chapter, several additional loci are described that appear, with *crp*136 and *crp*65, to comprise a '*vsp136*' subfamily of genes. These findings are novel and constitute a substantial contribution to the knowledge base on *Giardia* vsp genes. The significant findings described in this chapter are as follows:

Several vsp genes that contain tandem repeat elements were characterised in the Ad-1/c3 isolate of *G. intestinalis*. This isolate is distinct from the WB isolate but belongs to the same genetic subtype (A-I). The complete coding sequences of three loci, *vsp*136-2, *vsp*136-3 and *vsp*136-4, were characterised in this study using cloned genomic DNA fragments. Partial sequence data were also obtained for two other *vsp*136-like loci, *vsp*R1 and *vsp*R2, which were identified in PCR-amplified cDNA. These loci proved to be closely related to the *crp*136 gene described by Chen *et al.* (1995), with *vsp*136-2, *vsp*136-3 and *vsp*136-4 possessing 5' non-repeat coding sequences and tandem repeat elements that were identical with the corresponding segments of *crp*136. The near-identity of these multiple loci indicates that they have arisen by successive gene (or chromosomal) duplications and that they therefore represent a '*vsp*136' (*crp*136-like) gene subfamily. It seems very likely that *vsp*136-2, identified in the Ad-1 genome in this current study, is in fact the *vsp*136-1 (*crp*136) locus identified previously in the WB isolate by Chen *et al.* (1995), rather than a separate locus. This is evident from the almost 100% nt sequence identity that is apparent between these two genes, not only within the coding sequences (which contain an identical number of repeat units, 23.5 copies) but also in the flanking non-coding regions (**Figs. 3.18, 3.19**). This near-identity is evident also in the dendrogram derived by phylogenetic analysis of the deduced C-terminal amino acid sequences (**Fig. 3.17**). However, with the detection of other closely related loci, e.g. *vsp*136-3 and *vsp*136-4, in separate genomic DNA fragments, it would

require a full and detailed analysis of all *vsp*136-like genes that exist within the genome to unambiguously resolve this point.

Nucleotide sequence data from *vsp*136-2, *vsp*136-3, *vsp*136-4, *vsp*R1 and *vsp*R2, totalling 14,823 bp of new sequence, were obtained in this current project. These have been lodged with GenBank. The findings in this chapter, specifically differences in the number of copies of an invariant tandem repeat element in *vsp*136-2 (23.5 copies), *vsp*136-3 (1.5 copies) and *vsp*136-4 (20.5 copies) as well as the identification of highly conserved flanking sequences between the various characterised members of the *vsp*136 gene subfamily (including *vsp*52 and *crp*65), suggest that a series of duplication and recombination events has given rise to this gene subfamily. As was discussed in section 3.11, alignment of the flanking sequences from these loci indicated that several recombinations may have occurred between different segments (3' non-repeat and/or tandem repeat segments). To a large extent, the apparent conservation of nucleotide sequence between these loci may be explained if the serial duplications and recombinations that created these multiple loci occurred relatively recently. In this case, the similarities would not reflect conservation due to selective constraints but rather insufficient time for more than a few rare mutations to occur. Given a longer period of time, one might expect many more mutations to accumulate and cause a more substantial divergence between these loci. Alternatively, if the duplications are more ancient, mutations within the coding sequences might be expected to exhibit a bias toward synonymous (silent) sites or to conservative amino acid substitutions as a result of selection at the protein level (cells with deleterious mutations would be disadvantaged). The coding sequences do not show many mutations, suggesting that the gene duplications that gave rise to *vsp*136-2, *vsp*136-3 and *vsp*136–4 occurred relatively recently.

This project was initially begun with the aim of characterising functional *vsp*52-related genes in the Ad-1/c3 *G. intestinalis* genome. It succeeded in identify three new genes that are related more closely to *vsp*136-1 (*crp*136) than to *vsp*52. These loci were identified

by Southern hybridisation analyses of genomic DNA restriction fragments and they were isolated by screening genomic DNA libraries with probes prepared from segments of *vsp*R2 cDNA that had been amplified by PCR using primers designed from the *vsp*52 sequence. Although the *vsp*R2 locus was not identified among the cloned *Sac* I genomic fragments isolated from the genomic libraries, there are identifiable reasons that might explain this failure. The number of tandem repeats present in the *vsp*136-related genomic fragments was found to strongly influence the staining intensity of the bacterial colonies that harboured these plasmid constructs. This is exemplified in **Fig. 3.9**, in which constructs pM24-1 and pM42-2 were compared by Southern hybridisation. The segment of the *vsp*R2 transcript that was identified from Ad-1/c3 cDNA contained one full and two adjacent (flanking) half copies of a 120-bp repeat unit. If the *vsp*R2 gene possesses only a few tandem copies of this repeat, it is possible that constructs containing *vsp*R2 would have hybridised more weakly, i.e. stained more faintly (e.g. **Fig. 4.9**), than those containing other (*vsp*136 subfamily) loci which possess large numbers of a very similar repeat. Moreover, if recombinations occur at a high rate in these vsp genes (those containing tandem repeat sequences), it is conceivable that the *vsp*R2 locus in the majority of Ad-1/c3 trophozoites (and therefore, bulk genomic DNA) might have lost most of its tandem repeat region or had it replaced by a different repeat element. These issues cannot be clarified without identifying and characterising all of the related genes that exist within the genome.

The *Giardia* genome database is proving to be an extremely useful resource for the identification of new genes. As the random sequences are compiled into longer contiguous (chromosomal) sequences, the value of this resource will become even more significant. Although the assembly of single- or low-copy-number divergent sequences should present no real difficulties to the compilers, the correct compilation of multiple similar or identical sequences that form part of a large, complex gene family such as that evident from the vsp genes may not prove to be an easy task. Indeed, from the data presented in this chapter and in

the next chapter, the unambiguous compilation of some regions of the genome that contain many similar, replicated loci may require the physical isolation of various relevant genomic fragments, their definition by size and analysis by restriction mapping and nt sequence determination, as described in this thesis.

After identifying the *vsp*136-2, *vsp*136-3 and *vsp*136-4 genes in the Ad-1/c3 genome, the *Giardia* genome database was searched for the presence of *vsp*136-1 (*crp*136) related sequences using the 120-bp repeat of *vsp*136-1 and *vspR2*. Several unedited sequences were obtained, all of which contained segments identical to the sequence of the 120-bp *crp*136 repeat unit (data not shown). Although these sequences were 800-1200 bp long, they did not span the entire tandem repeat region. For this reason, and because of uncertainties arising from the unedited nature of these single sequence runs, it was impossible to compile them and to determine how many tandem repeats were missing or which specific 3' or 5' non-repeat sequence each belonged. As a result, it was impossible to identify with any certainty which particular gene each sequence represented, except for one sequence that appeared, on the basis of its 5' untranslated and 5' coding sequence, to represent the *vsp*136-1 (*crp*136) locus.

Chapter 4

The '*vsp*72' gene subfamily

The '*vsp*72' gene subfamily

4.1 The *vsp*72 gene subfamily: Introduction

In chapter 3, a preliminary Southern hybridisation analysis of *Sac* I restricted Ad-1/c3 *G. intestinalis* DNA with the non-repeat 3' segment of *vsp*R2 (*vsp*R2-[3']) as the probe revealed at least 8-10 related fragments (**Fig. 3.3B**). In comparison, only 3-4 fragments hybridised with the *vsp*R2-R probe, indicating that the genome contained a greater number of sequences related to the 3' non-repeat region of the *vsp*R2 insert than it did of sequences similar to the repeat region of the same *vsp*R2 insert (**Fig. 3.3**). In choosing to investigate the identity and character of these genomic *vsp*R2-[3']-like sequences, two experimental strategies were adopted:

1. To screen genomic libraries for hybridisation with the *vsp*R2-[3'] probe, in order to identify and clone these related fragments.

2. To amplify related sequences from restricted genomic DNA using the 3'-overhang extension method (described in sections 2.7.8 and 2.7.9). This technique is less stringent than conventional PCR, since it employs only one specific primer (in combination with the 3'-overhang extension primer). This enables related, heterogenous sequences (segments situated between the gene-specific primer site and a nearby restriction site cleaved by a chosen endonuclease) to be amplified in addition to the specific, targeted sequence. The method has proved useful for isolating and cloning gene segments, which have then been used to identify and clone particular loci for more rigorous analysis (Ey & Darby, 1998; Ey *et al.* 1999). However, the technique has limitations, e.g. in normal PCR it yields amplification products only if the DNA is cleaved by the selected endonuclease within 2-3 kb of the gene-specific primer site. It was considered an appropriate approach for investigating whether the additional genomic DNA fragments that hybridised with the

*vsp*R2-[3'] probe on Southern blots represented functional loci belonging to definable vsp gene subfamilies or non-functional (e.g. corrupted) loci, perhaps derived from previously functional genes.

With the aims of obtaining preliminary information about *vsp*R2-[3']-like sequences, it was decided to begin an analysis of genomic DNA from the Ad-1/c3 *G. intestinalis* clone using the 3' overhang extension method.

4.2 Detection of gene segments with similarity to the *vsp*R2-[3'] probe

The 3'-overhang extension technique is described in detail in sections 2.7.8 and 2.7.9. Three generic oligonucleotides (969, 970 and 109) were designed as templates for extension from the 3' overhanging ends of DNA cleaved by the restriction endonucleases *Kpn* I, *Pst* I and *Sph* I respectively. Each of these oligonucleotides (henceforth termed *Kpn* I, *Pst* I or *Sph* I 'primers') was used to extend the 3' ends of appropriately cleaved genomic DNA fragments and then tested in PCR in combination with oligo 121, the consensus reverse primer designed from alignment of the 3' segments of *crp*136 and *vsp*R2 (Fig. 3.2), for their ability to amplify sequences related to *vsp*R2. PCR controls included tubes lacking template DNA or either primer (section 2.7.9). Comparison of the products of test reactions with any product(s) obtained in the controls indicated whether all (or which) products amplified in the test reactions required both primers for their synthesis.

Using oligo 121 in combination with the *Kpn* I, *Pst* I or *Sph* I primers, a variety of amplification products was obtained with a size range of 300-2000 bp as exemplified in **Fig. 4.1A**. To examine the nature of the amplified DNA and to determine which might warrant further analysis, they were tested for hybridisation on Southern blots using the *vsp*R2-[3'] probe (**Fig. 4.1B**). Of the products amplified using the *Sph* I primer + oligo 121 (not shown), those which hybridised with the probe were relatively short (300-500 bp) whereas other

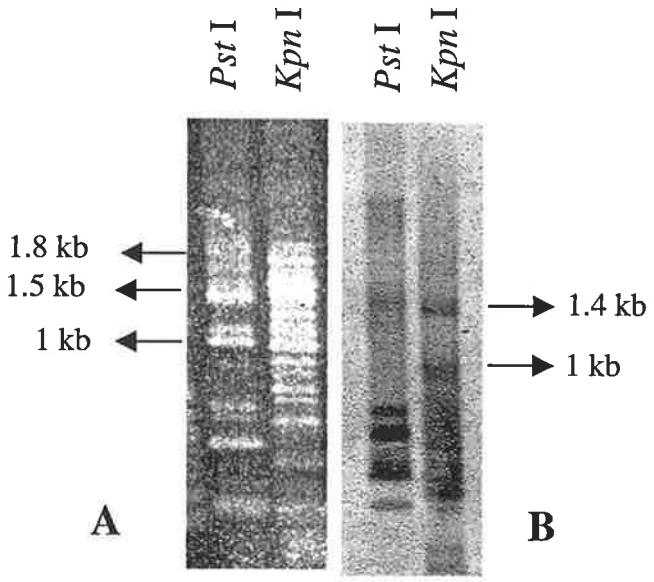


Figure 4.1. PCR amplification of restricted *Giardia* genomic DNA (3' extended using '*Pst* I' or '*Kpn* I' primer templates) and amplified in combination with oligo-121 (see Fig. 4.3).

(A) Amplified products were separated by electrophoresis and stained with ethidium bromide.

(B) Southern hybridisation of the same gel (A) with the *vsp*R2-[3'] probe.

products, e.g. those amplified using the *Kpn* I or *Pst* I primers in combination with oligo 121 ranged from 300-1500 bp (**Fig. 4.1 B**). Several of the bands were prominent on the Southern blots.

The DNA in these PCR reaction mixtures was subsequently blunt-ended using T4 DNA polymerase and cloned into pGEM-3Zf (+) (section 2.10.2). Several clones were chosen and plasmid DNA was extracted from the overnight growth of bacterial cultures. The plasmid inserts were subsequently tested for hybridisation with the *vsp*R2-[3] probe using linearised plasmid DNA. Constructs containing inserts that hybridised with the *vsp*R2-[3'] probe are shown in **Fig. 4.2**. Several clones were chosen for further analysis. Two of these, pM20 and pM21 (**Fig. 4.2**), contained 1.3-kb and 1-kb inserts respectively that had been amplified from *Kpn* I-restricted genomic DNA in PCR using the *Kpn* I primer and oligo 121. Two other constructs (pM30 and pM41) harboured inserts of 1.3 kb and 1.48 kb respectively (**Fig. 4.2**) that had been amplified using the *Pst* I primer and oligo 121 (**Fig. 4.3**).

These four inserts were sequenced, initially by using T3 and T7 primers that hybridise with the promoters that flank the multiple cloning site of the vector. Additional primers were also designed (as shown in **Fig. 4.3**) and used to determine the complete sequence of each insert. Alignments of the pM21 and pM20 insert sequences showed that they were almost identical (99% identity over a 1 kb overlap, Table 4.1). Each insert represented an uninterrupted segment of a longer reading frame that appeared to be part of a functional vsp gene, based on the apparent integrity of the coding sequences. An alignment of the inferred amino acid sequences (**Fig. 4.4**) revealed that these shared 98% amino acid identity over the 340 residues that were encoded by the cloned inserts.

An interesting feature evident from these alignments was that polymorphisms in the *Kpn* I and *Pst* I restriction sites were responsible for their differential amplification (using the *Kpn* I or *Pst* I primers in combination with primer 121) and isolation as similar but different-sized cloned amplification products that possessed identical 3' ends fixed by the use of

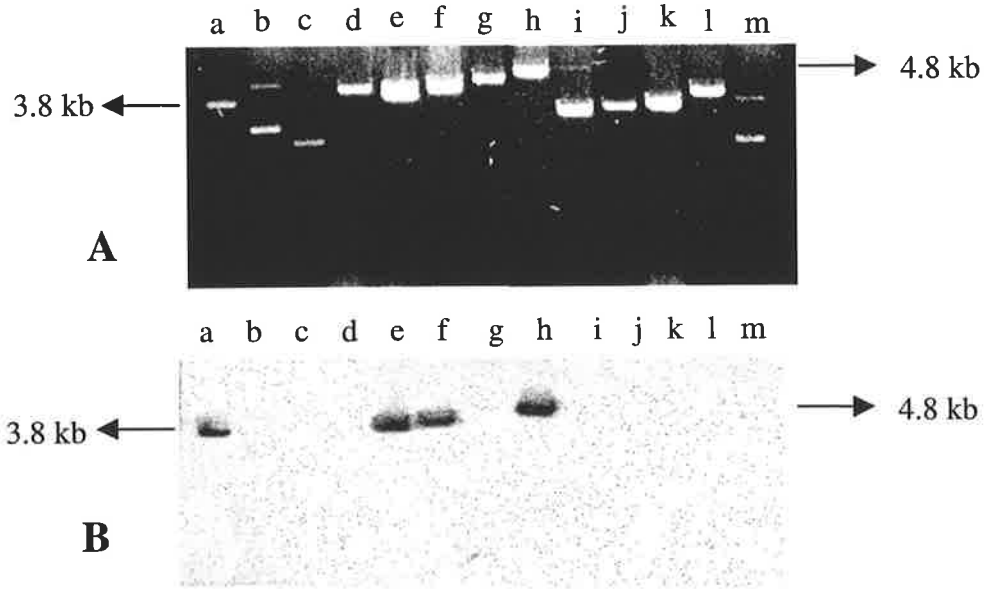


Figure 4.2. Restriction analysis of cloned plasmid constructs containing inserts derived by PCR amplification from template *Giardia* DNA that had been restricted with *Kpn* I and subjected to 3'-overhang extension using the *Kpn* I 'primer', oligo 969.

(**A**) Cloned constructs, linearised with *Xba* I, electrophoresed on a 1% agarose gel and stained with ethidium bromide.

(**B**) Southern blot of the same gel from (A) hybridised with the DIG-labelled *vsp*R2-[3'] probe. Lanes a, e, f, and l corresponded to constructs p*vsp*R2, pM21, pM20 and pM18 respectively.

Figure. 4.3. Segments of four VSP genes, identified from Ad-1/c3 *G. intestinalis* genome. Bold purple and red lines represent segments amplified using oligos 154 + 155 (M165-[5']) or 148 + 149 (M165-[3']) respectively. The later segment is also present in M20 and M21 sequences. These amplified products were used as probes for hybridisation analysis. Diagnostic restriction sites are indicated and PCR primer sites are depicted as arrows.

Cloned
PCR Product

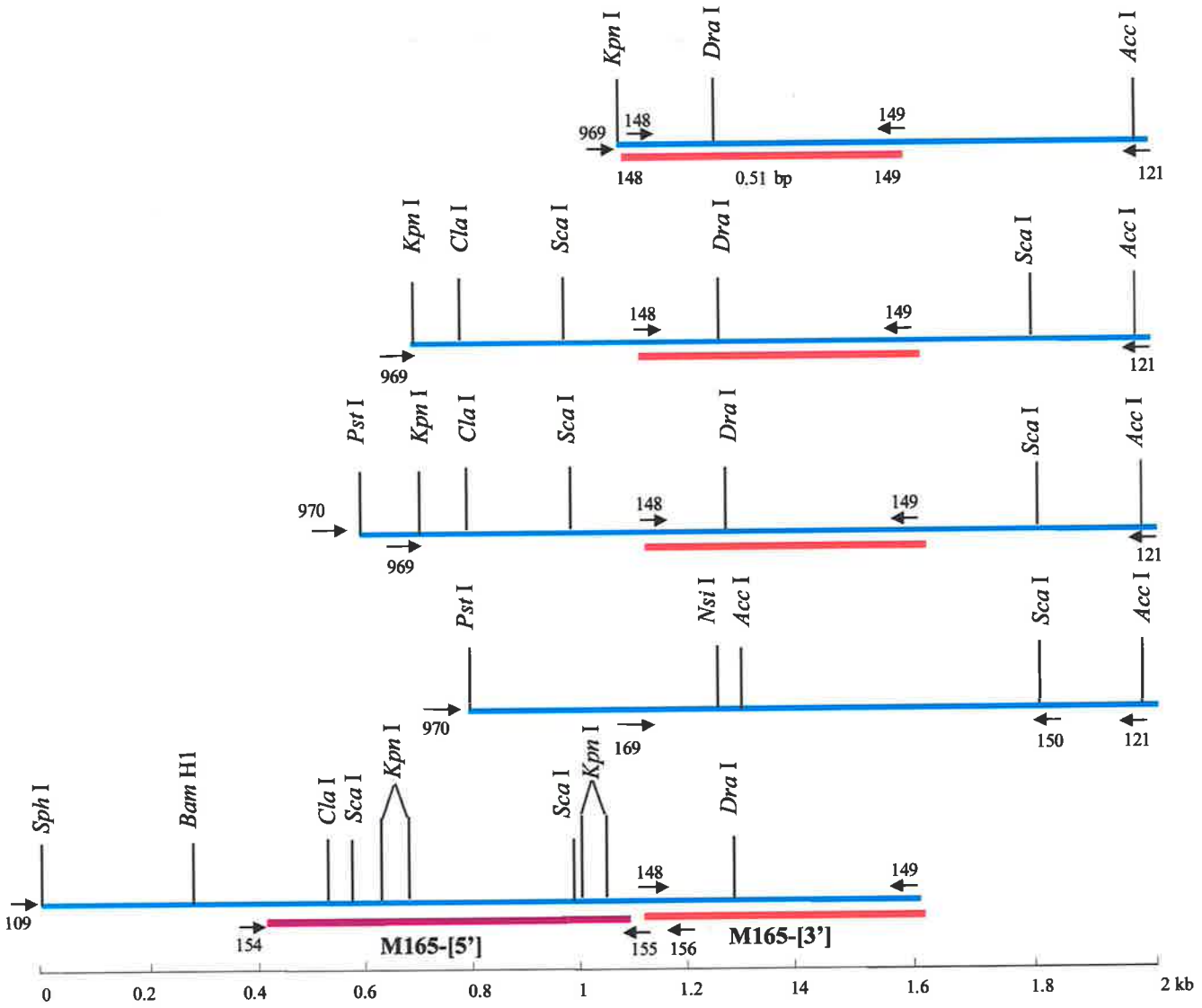
M21

M20

M41

M30

M165



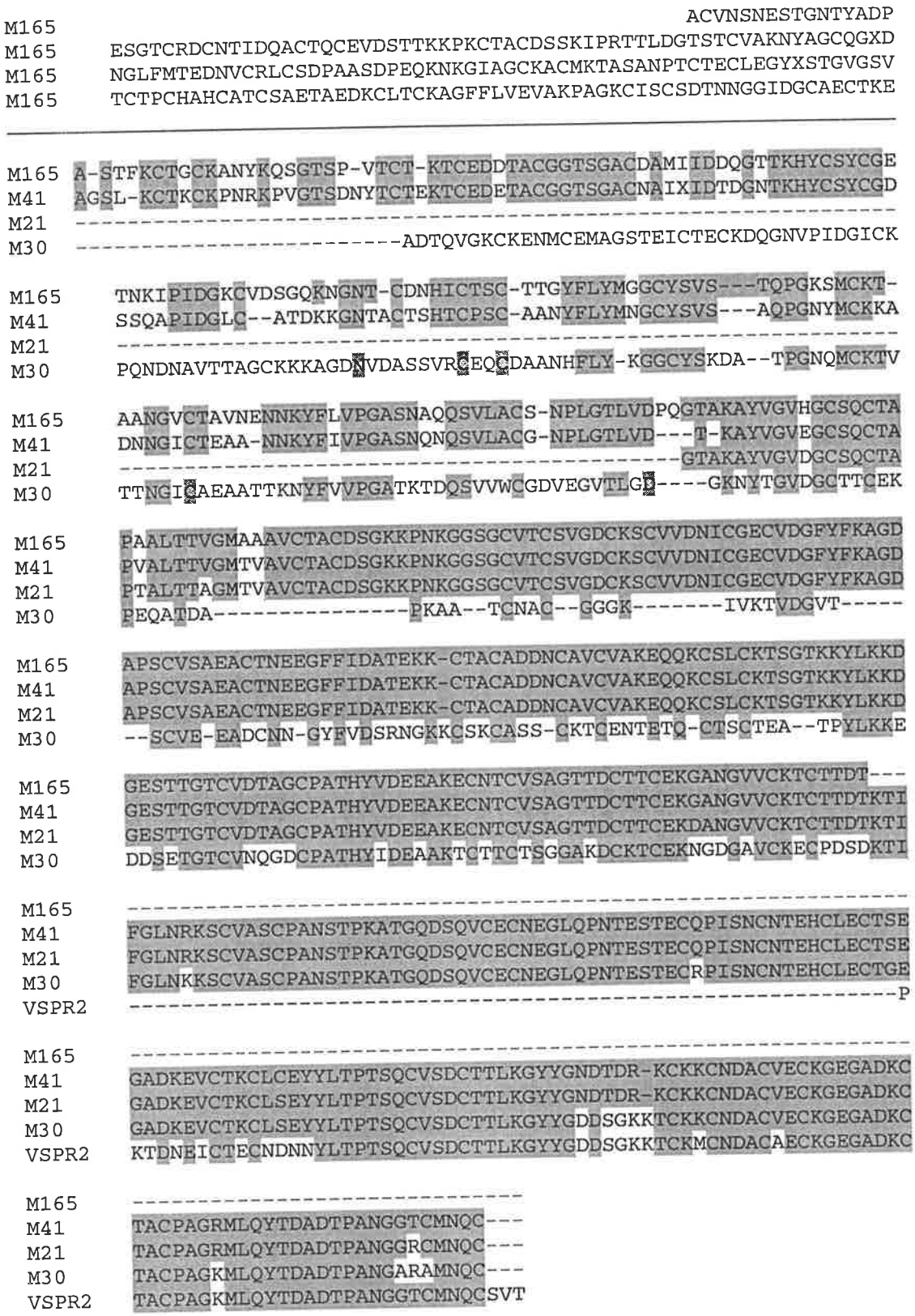


Figure 4.4. Alignment of VSP amino acid sequences M165, M41, M21 and M30 encoded by amplified segments of genomic DNA. The overlapping (homologous) segment encoded by *vsp*R2 (Fig. 3.2) is included in the bottom blocks of the alignment. Identical residues are shaded grey. Gaps (-) introduced by CLUSTAL are shown within each sequence. Positions for which no data were available are shown by red dashes. The following sequences are available under the gene bank accession numbers as indicated: M165, AF298863; M41, AF298864; M21, AF298865; M30, AF298866.

primer 121 (**Fig.4.3**). For example, the *Kpn* I site at the 5' end of the pM21 insert is missing from the pM20, pM41 and pM30 inserts but it is present in the pM165 insert, which was amplified from an *Sph* I fragment using the *Sph* I primer (oligo 109) in combination with primer 149 (**Fig. 4.3**). Unlike the pM20 and pM41 inserts, which were both amplified from *Pst* I genomic fragments using the *Pst* I primer (oligo 970), the pM20 insert was amplified from a *Kpn* I genomic fragment using the *Kpn* I primer (oligo 969). The latter insert was distinguished from the shorter pM21 insert because it lacked the *Kpn* I site at which the pM21 sequence was cleaved and subsequently amplified (**Fig. 4.3**). Similarly, despite the similarity of their sequences (evident from **Fig. 4.4**) the *Pst* I site of pM30 was absent (polymorphic) in the pM20, pM41 and pM165 inserts. However, inserts M21, M20, M41 and M30 all had common 3' sequences - due to the common 121 primer acting as a 'locus'-specific 3' PCR 'anchor' (**Fig. 4.3**).

All of these inserts also contained a common central sequence from which primers 148 and 149 were designed. The pM165 insert was subsequently amplified (described in section 4.3) using primer 149 as the 'subfamily'-specific 3' anchor, in combination with the *Sph* I primer (oligo 109). It was recovered after colony hybridisation using a probe prepared from template DNA produced in PCR using primers 148 + 149. The 5' *Kpn* I site of insert M20 was evident as an internal site in both the pM41 and pM165 inserts (**Fig. 4.3**). These comparative results indicated clearly that these different amplification products were indeed closely related but polymorphic loci. Complete sequencing of the pM20 and pM41 inserts revealed that they were identical over the entire segment that corresponded to the 1,372-bp pM20 insert (**Table 4.1**). The pM41 insert had an additional 116 bp at its 5' end, which extended from the internal *Kpn* I site to the 5' *Pst* I site used for its amplification and isolation. The identity of the pM20 and pM41 inserts suggested that they had been amplified, using the *Kpn* I and *Pst* I primers, respectively, from distinct *Kpn* I and *Pst* I genomic restriction fragments that overlapped and encompassed the same vsp gene. The pM21, pM30

Table 4.1. Similarity between PCR products amplified from 3′-overhang extentend genomic template DNA and the *vsp*R2-[3'] region.

	Cloned insert and length					
	MM20 1372-bp	MM21 1027-bp	MM30 1309-bp	MM41 1488-bp	MM165 1637-bp	vspR2 275-bp
	% Nucleotide identity (length of overlap, bp)					
pM20	-	99 [1020]	64 [1331]	100 [1372]	86 [945]	85 [258]
pM21		-	71 [1008]	99 [1020]	98 [581]	85 [258]
pM30			-	64 [1331]	51 [891]	88 [247]
pM41				-	84 [1058]	85 [258]
pM165					-	50 [255]

Percentage of nucleotide sequence identity determined by FASTA, as implemented in DNASIS. Numbers in parentheses indicate the length of the overlaps (bp) for which the % identity occurs.

and pM165 inserts appeared to be derived from distinct loci, as they were different from each other as well as from the pM20/pM41 sequence (**Fig. 4.3**). The amino acid sequence encoded by the pM30 insert exhibited 65.6% and 66% sequence identity over 302 residues with the polypeptides encoded by the pM41 and pM21 respectively (data not shown). However the level of nucleotide sequence identity was significantly higher (82%) near the 3′ ends of the cloned inserts. **Table 4.1** summarises the level of nucleotide sequence identity that was found between these various cloned amplification products and the *vsp*R2-[3'] sequence.

4.3 Attempts to obtain the flanking sequences of the pM20, pM30 and pM44 inserts

To obtain the 5' ends of the putative vsp genes represented by two of the amplified segments (pM21 and pM41 inserts), reverse PCR primer (oligo 149, corresponding to a conserved sequence in the centre of the pM21 insert but nearer the 3' end of the longer pM41 insert, **Fig. 4.3**) was designed. As mentioned in the previous section, this was used for PCR in combination with the *Sph* I primer. *Kpn* I and *Pst* I cleavage sites had been found in some of the previously characterised amplified products e.g pM21 and pM41 plasmid inserts (**Fig. 4.3**). This yielded a number of additional, new amplification products (not shown). To determine whether any of these amplification products was related to the pM21 or pM41 plasmid inserts, Southern hybridisations were performed using a 500-bp DIG-labelled probe (M20-[3']), that corresponded to the 3' portion of the pM20 plasmid insert. This probe was synthesised using template DNA amplified from pM20 in PCR using primers 148 and 149 (**Fig. 4.3**). Several of the new (primer 149-derived) amplification products hybridised strongly with this M20-[3'] probe. One strongly-hybridising 1.6-kb product was gel-purified and cloned into pGEM-3Zf (+) for additional analysis. The 1.6-kb cloned insert from one of these constructs (pM165) was subjected to nucleotide sequence analysis. As shown in

Fig. 4.3, this 1.6-kb DNA represented a segment of another putative vsp gene, related to (but distinct from) the gene segments identified previously in pM20, pM21, pM30 and pM41. However, the pM165 insert was more closely related to the pM20 and pM21 inserts (86% and 98% nt identity over 945 and 581 bp, respectively) than to the pM30 insert (51% nt identity over 891 bp) (**Table 4.1**). Conserved restriction sites (for *Kpn* I, *Sca* I and *Dra* I) were identified between the pM41 and pM165 insert sequences, as shown in **Fig. 4.3**. More attempts were made to obtain the flanking sequences of the pM21, pM30, pM41 and pM165 inserts by the 3'-overhang extension method using oligos 148, 150 and 169, each in combination with the *Kpn* I, *Pst* I or *Sph* I primers as described earlier (section 2.7.9). Several amplified DNA products were obtained for each primer combination. Some of these products were cloned and partially sequenced, revealing that they were segments of highly homologous loci (sharing 98-99% nt identity over their entire lengths [1200-1500-bp] with pM30, pM41 or pM165). At this stage, it was evident that attempts to identify and assemble amplified sequences from the different members of a complex gene family that contained such highly homologous loci (many of which appeared to be near-identical genes) would be a futile exercise. The second approach, to construct and screen *Giardia* genomic DNA libraries for loci related to these characterised gene segments (pM21, pM30, pM41 and pM165), was therefore commenced.

4.4 Similarity with known vsp genes

Nucleotide and inferred amino acid sequence alignments between the pM20 and pM21 inserts and published vsp gene sequences (data not presented) showed that the inserts had greatest similarity with *crp*72, a vsp gene encoding a 68-kDa polypeptide that was described by Adam *et al.* (1992). At the nucleotide level, *crp*72 exhibited 65% sequence identity with the pM21 insert over a 390-bp overlap, 62% identity with the pM41 insert over 891 bp, and 66% identity with the pM165 insert over 1464 bp.

To summarise these results:

a) Use of the 3'-overhang extension technique led to the amplification (from restricted *G. intestinalis* Ad-1/c3 DNA) and identification of multiple DNA products that appeared to be internal segments of closely-related vsp genes. Efforts to obtain the complete genes represented by these particular inserts were not successful, as additional amplification products were derived from other closely related loci.

b) None of the amplified segments contained tandem repeats. This was in contrast to the *vsp*136 subfamily loci described in chapter 3 and somewhat surprising, since oligo 121 (used in the amplification of these gene segments), was a reverse (antisense) primer designed to hybridise to a common sequence within the non-repeat 3' end of the *crp*136 and *vsp*R2 coding sequences (**Fig. 3.2**). However, use of this antisense primer in combination with the *Pst* I, *Kpn* I or *Sph* I primers on the appropriately restricted, 3'-overhang extended template DNA resulted in amplification of vsp gene segments which are similar within this 3' region (over ≈ 270 bp) with *vsp*R2; (see **Table 4.1 and Fig. 4.4**) but appear to belong to a separate vsp gene subfamily whose members lack tandem repeats.

c) The failure to identify segments of *vsp*136-like loci among the 3'-overhang extended genomic fragments amplified using oligo 121 as the reverse primer may have been due to a lack of the necessary proximal restriction sites, i.e. no site(s) may exist within a sufficiently close distance of the oligo 121 priming site for successful PCR amplification of these particular genes. However, the presence of *vsp*136 subfamily-derived sequences within the amplified DNA cannot be excluded, since few of the amplified products were cloned and characterised.

d) Although the 3'-overhang extension method has unique potential (in this case, the capacity to amplify segments from a number of related loci which share the oligo 121

nucleotide sequence), only those segments of each locus that contained an appropriate *Kpn* I, *Sph* I and/or *Pst* I site were amplified. Thus, segments that could be amplified from 3'-overhang extended DNA probably represent only a fraction of related loci. Evidence from these experiments for the existence of multiple similar sequences made it clear that the only way to characterise unambiguously the genes that are represented by (or related to) the pM20, 21, 44 and 165 inserts was to isolate and characterise larger genomic fragments by screening genomic libraries.

In summary, analysis of these five clones (pM20, pM21, pM30, pM41, and pM165) led to the identification of segments of what appear to represent a large vsp gene subfamily. These loci appeared to have 3' segments similar to the 3' segment of *vsp*R2 (**Table 4.1, Fig. 4.4**), but they did not appear, from the sequence data obtained, to contain tandem repeat sequences.

4.5 Identification of loci related to the pM165 insert

To identify the pM165 locus and other related loci within the *G. intestinalis* genome, DNA from the Ad-1/c3 isolate was subjected to cleavage by various restriction endonucleases and then analysed by Southern hybridisation, using a probe (M165-[5']) prepared from a 686-bp amplified segment (primers 154 + 155) in the 5' region of the pM165 insert (**Fig. 4.3**). The result, illustrated in **Fig. 4.5** , indicated the existence of numerous (at least 10-15 fragments) that stained strongly. This indicated that the genome contained numerous sequences related to the 5' segment of this (pM165) sequence.

4.6 Identification of a novel vsp gene subfamily

With the aim of elucidating the nature of the loci represented by the pM41 and pM165 inserts, i.e. whether they represented a subfamily of divergent and functional vsp

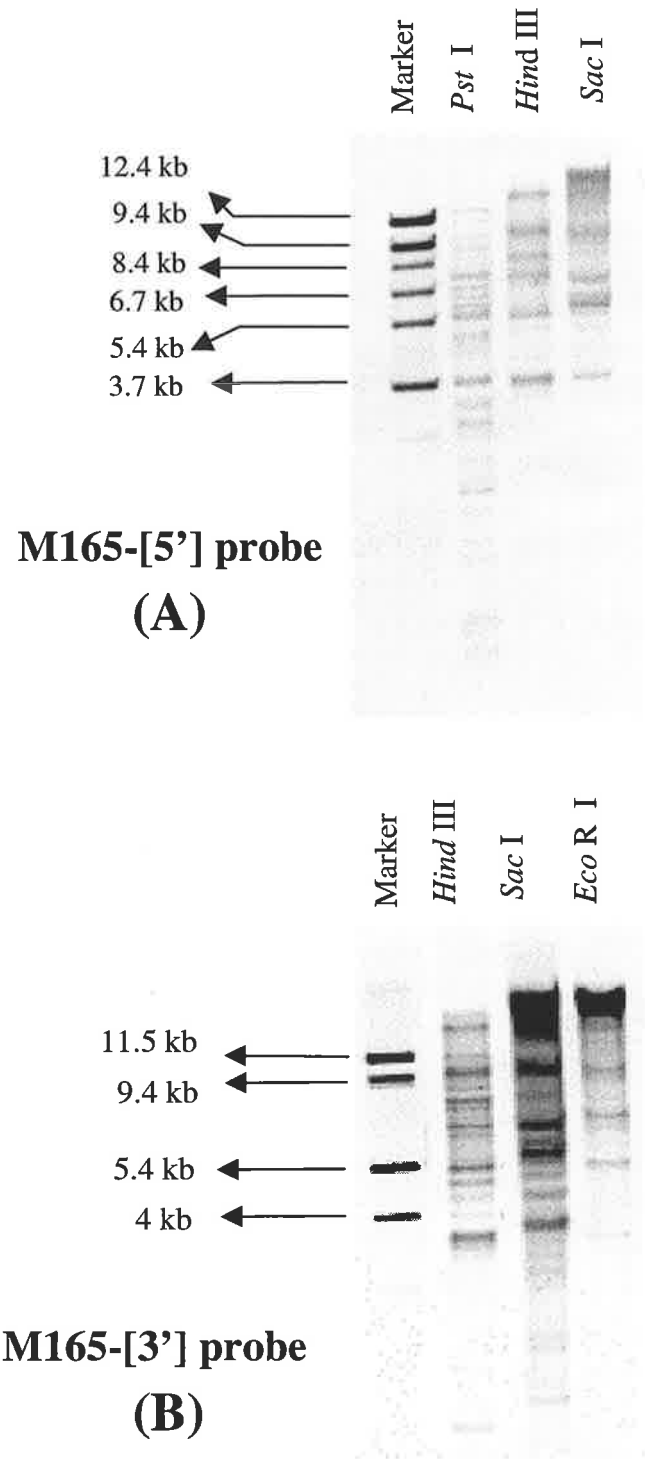


Figure 4.5. Southern hybridisation analysis of the *G. intestinalis* genome. Chromosomal DNA from *Giardia* (Ad-1/c3 clone) was restricted by incubation with (A) *Pst* I, *Hin*d III or *Sac* I, or (B) *Hin*d III, *Sac* I or *Eco* RI and tested for hybridisation with probes derived from the 5' (A) or 3' (B) segment of the pM165 insert. The sizes of the DIG-labelled markers are indicated by arrows.

genes or merely corrupted (pseudo) genes, a decision was made to clone and characterise some of these fragments. To this end, several genomic libraries were constructed in the plasmid vector Bluescript-SK(+) from *Sac* I restricted Ad-1/c3 *G. intestinalis* DNA. Segments of the pM165 insert were used to make DIG-labelled probes for screening these libraries, as depicted in **Fig. 4.3**:

a) A 686-bp segment from the 5' region of the insert, amplified using oligos 154 + 155.

b) A 500-bp segment from the 3' end of the insert, amplified using oligos 148 + 149.

It was assumed that each probe might hybridise with a different subset of vsp gene sequences, although some common fragments, with greater similarity to the pM165 plasmid insert, hybridised with both probes.

Three independent genomic libraries were constructed from *Sac* I-restricted Ad-1/c3 *G. intestinalis* DNA. Transformants from each library were screened by colony hybridisation using the aforementioned probes and a total of 20 clones were identified for investigation, as exemplified in **Fig. 4.6**. These were grown in liquid culture and the plasmid DNA was extracted and purified for analysis. Some of the colonies required recloning to ensure they contained only the plasmids of interest.

4.7 General characterisation of the cloned *Sac* I restriction fragments

All 20 plasmid constructs were incubated with *Sac* I and subjected to electrophoresis to determine the approximate sizes of the inserts. These varied substantially, ranging from 3.6 to 13 kb, with very few clones appearing to be replicates, i.e. possessing an insert of a size similar to that in any of the other clones. All were also subjected to Southern hybridisation analysis (using *Sac* I-cleaved plasmid DNA) to confirm that all of the inserts had similarity to the pM165 insert.

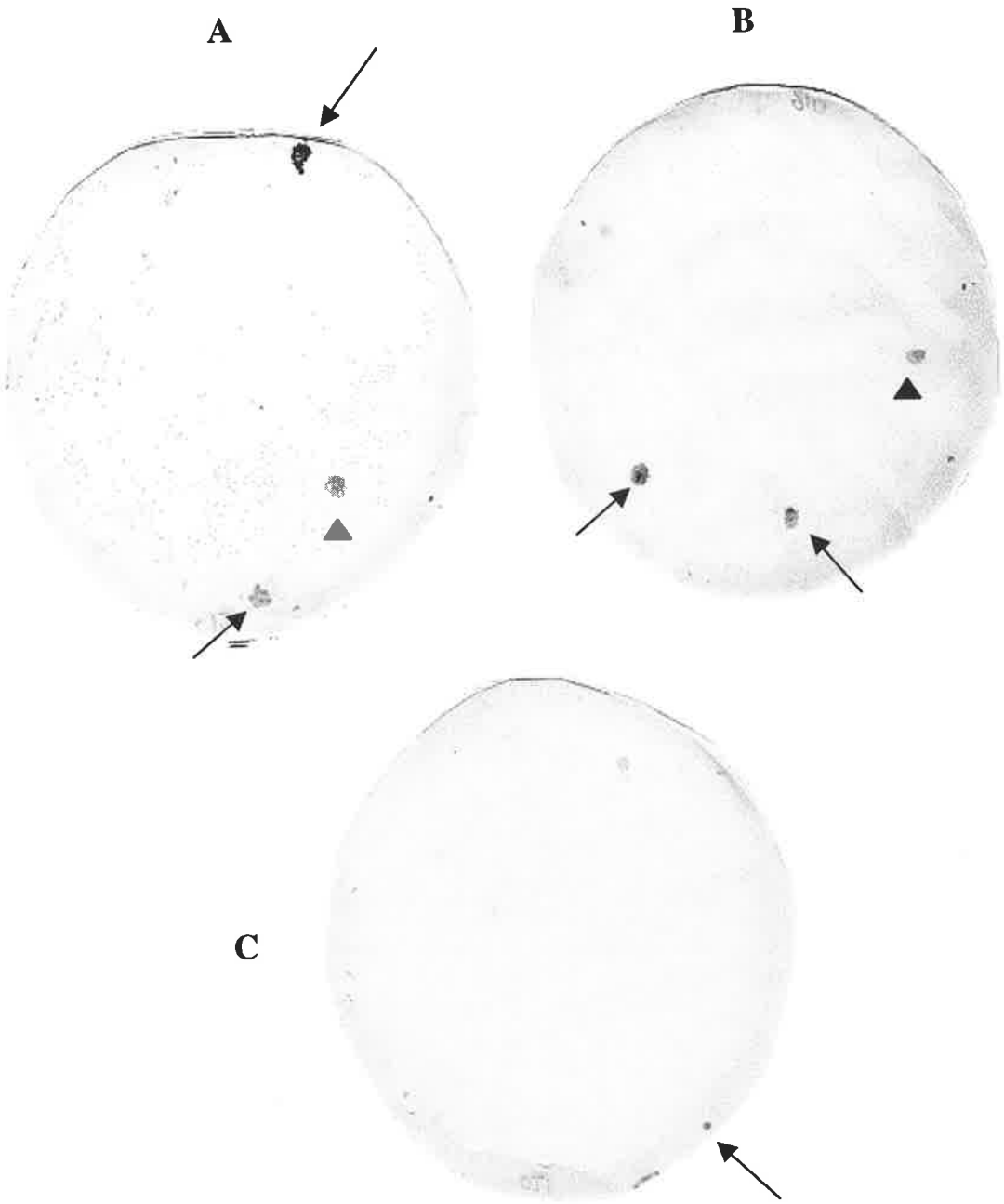


Figure 4.6. Detection of colony blot hybridisation.
Colonies representing part of a genomic library generated from *Sac* I fragments of *G. intestinalis* DNA were screened for hybridisation with the DIG-labelled **M165-[5'] probe**. Arrows show clones that hybridised with the probe. An arrowhead shows the M165 clone (positive control). Each plate contained 200-500 colonies.

Table 4.2. List of the 20 plasmid constructs containing cloned *Sac* I genomic DNA fragments that hybridised on colony blots with pM165 probes

	Plasmid identity		Size of	Probe	Results			Predicted VSP
List #	Clone number	1-Letter code	insert (kb)	used for S/Hyb.	Southern hybr.	Analytical PCR	Nucleotide sequencing	gene sequences (tandem arrays)
1	11-2	(A = S)	5.43	A[a]	+	+	+	+
2	11-3	(B)	5.5	A	+	+	+	+
3	7-1	(C = R)	9.5	A	+	+	+	+
4	13-3	(D)	3.76	A	+	+	+	+
5	3-3	(E)	8.45	B[b]	+	+	+	+
6	5-1	(F)	6.8	B	-	+	-	Not examined
7	9-1	(G)	4.6	B	-	+	-	Not examined
8	12-8	(H)	9	A	-	+	-	Not examined
9	14-3	(I)	3.6	A	-	+	-	Not examined
10	16-1	(J)	9	A	-	+	-	Not examined
11	16-3	(K)	12	A	-	+	-	Not examined
12	17-1	(L)	10	A	-	+	-	Not examined
13	17-4	(M)	9.5	A	-	+	-	Not examined
14	17-7	(N)	7.3	A	-	+	-	Not examined
15	17-9	(O)	6.6	A	+	+	-	+
16	19-2	(P)	5.9	A	-	+	-	Not examined
17	21-1	(Q)	13	A	-	+	+	Not examined
18	3-2i	(R = C)	9.5	B	+	+	+	+
19	3-1d	(S = A)	5.43	B	+	+	+	+
20	3-2b	(T)	7	B	+	+	+	+

[a] The M165-[5'] probe, prepared in PCR from pM165 template DNA using primers 154 + 155 (see Fig. 4.3)

[b] The M165-[3'] probe, prepared in PCR from pM165 template DNA using primers 148 + 149 (see Fig. 4.3).

All 20 clones were then tested further by restriction analysis (using about 12 endonucleases) to construct maps of the inserts, by PCR amplification (described in detail in section 4.9) for 9 of the constructs, and by more detailed Southern hybridisations to locate the region(s) of interest within the inserts, some of which were very large (9-10 kb). **Table 4.2** shows the identity of the plasmids and a summary about how each was analysed.

No common restriction pattern was detected among the 20 inserts. However, two pairs of clones contained inserts that appeared to be identical in size and thus represented possible replicates (≈ 9.5-kb inserts in both pM7-1 and pM3-2I, and ≈ 5.43-kb inserts in both pM3-1D and pM11-2). These inserts were among those that were subsequently subjected to nucleotide sequence analysis, which indicated that they were identical but cloned in opposite orientations. Clones 11-2 and 7-1 were identified by hybridisation with the M165-[5'] probe, whereas clones 3-2I and 3-1D were identified by hybridisation with the M165-[3'] probe.

4.8 Detailed analysis of five genomic restriction fragments

4.8.1 General summary

It was neither sensible nor feasible in the time frame of this project to analyse and characterise all of the 20 cloned genomic fragments. However, five were chosen for a detailed, comparative study. Two of these (7-1 and 11-2) had hybridised to both the 5' and 3' M165 probes. The other three inserts (11-3, 13-3 and 3-3E) were chosen at random from the remaining 18 in the panel. All five constructs were examined by detailed restriction analysis using both single- and double-enzyme incubations to construct maps of the inserts, using size estimates (of the resulting restriction fragments) derived by regression analysis of fragment electrophoretic mobilities. Selected gels were additionally subjected to Southern hybridisation analysis, to identify the (internal) fragments of interest. In most cases, multiple fragments from each single enzyme digestion were found to hybridise with the DIG-labelled

M165-3' or 5' probes. Preliminary nucleotide sequencing was undertaken from both ends of the inserts using T3 or T7 primers, which hybridise to the respective promoters that flank the multiple cloning site of the vector. The resulting sequence data indicated that the inserts (the exception being pM3-3E) contained more than one segment that was related to the probe. **Fig. 4.7** shows a general outline of the open reading frames that were identified within these five genomic fragments. The preliminary nucleotide sequence data were used to design additional sequencing primers and the complete sequence of each insert was determined from overlapping runs using both strands. This required the construction of various subclones, because it quickly became apparent that each parent clone possessed multiple sites to which many of the primers hybridised. For example, in order to completely sequence the 5.44-kb insert of pM11-2, nine subclones were constructed (see **Fig. 4.13**). In compiling the final nucleotide sequences, each of the five inserts was found to contain a tandem array of vsp pseudo genes arranged in a head-to-tail arrangement (**Fig. 4.7**). Some of these coding sequences were 5' or 3' truncated, as they were located at the ends of the cloned insert(s) and had been cleaved at a *Sac* I site situated within an open reading frame. However, for those open reading frames that exhibited significant similarity with the M165 sequence the following common features were identified:

1. All were pseudo genes that lacked only the initial 5' end (≈300-bp) of functional vsp genes, i.e. a segment encoding the N-terminal methionine, signal peptide sequence and a short subsequent N-terminal portion. Otherwise, the coding sequences appeared intact and structurally similar to functional vsp genes.

2. The inferred amino acid sequences were typical of *Giardia* VSP, each possessing multiple copies of the 'CXXC' motif and a conserved hydrophobic (transmembrane) C-terminal segment that was followed by the invariant C-terminal VSP signature sequence, -CRGKA$_{COOH}$.

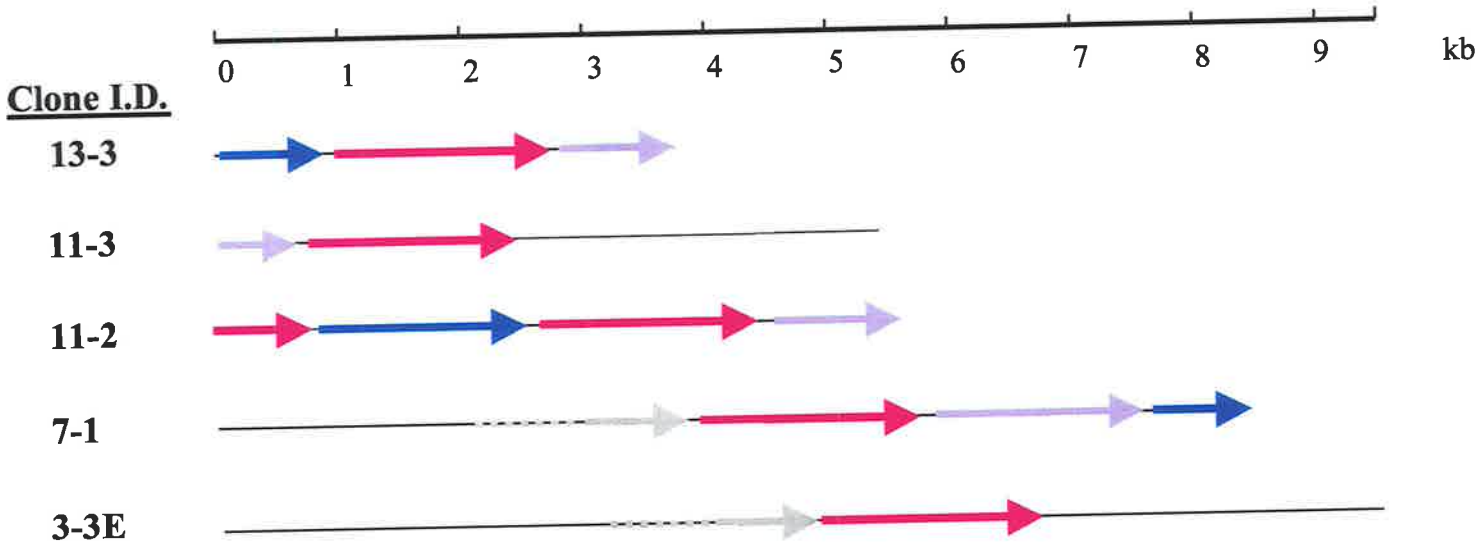


Figure 4.7. Linear depiction of the five genomic DNA inserts that were characterised by nucleotide sequence determination. The inserts are drawn to scale and identified open reading frames are indicated by bold arrows. Dashed segments represent the segments of open reading frames that were not completely sequenced.

3. Each pseudo gene possessed at its 3' end an uncorrupted polyadenylation signal sequence (AGTRAA) (Adam, 1991), situated in all cases 8-bp beyond the stop codon and preceded by the vsp gene specific 'spacer' motif 'CTTAGRTAGTRAA' (Svärd *et al.* 1998; Ey *et al.* 1999).

4. Each pseudo gene was separated from its neighbour by only a short intergenic segment.

5. A novel feature of the pseudo genes, not described previously in the literature, was their presence as a tandem gene array.

6. Most of the pseudo genes were highly related to the *vsp*1269 (*crp*72) gene described by Adam *et al.* (1992).

4.8.2 Detailed analyses

4.8.2a Clone C (pM7-1)

Two constructs (pM3-2I, pM7-1) were identified with similar sized (≈ 9.5-kb) *Sac* I genomic inserts. One construct (pM3-2I) had been identified by hybridisation with the M165-[3'] probe, the other (pM7-1) by hybridisation with the M165-5' probe (**Fig. 4.3**). Characterisation of the two inserts by restriction mapping (**Fig. 4.8**), Southern hybridisations (**Fig. 4.9**) and partial sequence determinations showed that they were identical but cloned in opposite orientations. The results of the restriction analyses of pM7-1 are depicted in **Figs. 4.8** and **4.9** and summarised in **Tables 4.3** & **4.4**. Subclones of pM7-1 were constructed from fragments generated by cleavage with *Bam* HI, *Pst* I or *Hin*d III. The fragments were cloned into pBluescript SK(+), generating 5 recombinant plasmids with inserts as illustrated in **Fig. 4.10**. Both strands of the subcloned restriction fragments were subjected to sequence analysis in order to compile the sequence of most of the parent insert. Five sequencing primers were designed for this purpose. These were used together with several other primers (**Fig. 4.10**) to

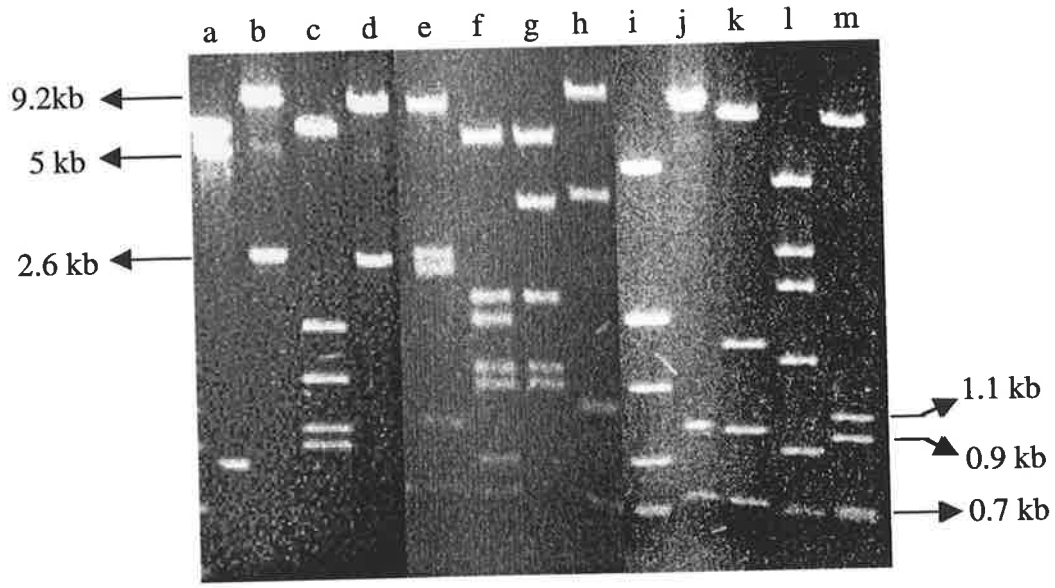


Figure 4.8. Double-enzyme restriction analysis of plasmid construct pM3-2i (7-1). Replicate aliquots of plasmid DNA were incubated overnight with different combinations of restriction endonuclease, as indicated.

a	*Acc* I	f	*Sph* I + *Xho* I	k	*Dra* I + *Hind* III
b	*Apa* I	g	*Sph* I + *Eco* R I	l	*Cla* I + *Xho*
c	*Bsp* H I	h	*Hind* III + *Xho* I	m	*Dra* I + *Xho* I
d	*Sma* I	i	*Bam* H I + *Sph*		
e	*Kpn* I + *Dra* I	j	*Dra* I + *Eco* R I		

Table 4.3. Single-enzyme restriction endonuclease cleavage data and associated Southern hybridisation data for plasmid construct pM7-1 (construct C)

Single - enzyme cleavage tests

Enzyme tested	No. of fragments	Sites in vector	Sites in insert	Fragment sizes (kb)	Sum of fragments (kb)	Identity* of fragment (site1-site2)	Hybridisation with probe 148-149
***Bam* H1**	4	1	3	4.80 3.80 2.70 0.90	12.2	B3 - Bo	- ++++ +/-
***Cla* I**	5	1	4	4.60 3.00 2.30 1.80 1.50	13.2	C2 - Vo Co - C1	+/- - +++ +++ +++
***Hin* d III**	2	1	1	12.00 0.72	12.7	H2 - Vo VoH - H2	
***Pst* I**	4	1	3	4.90 3.10 2.60 2.40	13.0	Po - P1 P3 - Vo P2 - P3 P1 - P2	++++ - - +
***Sph* I**	5	0	5	5.20 3.20 1.80 1.25 1.20	12.7	S5 - S1	++++ - + ++++ ++
***Xba* I**	4	1	3	6.10 3.50 1.80 1.10	12.5	Xba2-Xba3 Xba3 - Vo Xba1-Xba2 Xo - Xba1	++++ - +++ +/-
***Xho* I**	2	1	1	7.50 3.80	11.3	Xho - Xh1 Xh1 - Xho	++++ -
***Kpn* I**	3	1	2	9.00 2.10 1.30	12.4	K2 - Vo K1 - K2 Ko - K1	nt
***Sma* I**	2	1	1	7.30 6.20	13.5		nt

Average sum of fragments:	12.61	kb
***less* Size of vector:**	2.95	kb
Calculated size of cloned insert:	**9.66**	**kb**

* Cleavage sites are designated by the first letter of the relevant enzyme, followed by the order of the site(s) in the cloned insert, e.g. B_3 - B_0 represents the fragment spanning *Bam* H1 site (1) to the 3' end of the insert, which was cleaved at BamH1 site (0) within the multiple cloning site of the vector. Similarly, K_1 - K_2 represents the fragment spanning *Kpn* I sites (1) and (2).

Table 4.4. Restriction endonuclease cleavage data for plasmid construct pM7-1 (construct C)

Double - enzyme cleavage

Enzyme tested	No. of fragments	Sites in vector	Sites in insert	Fragment sizes (kb)	Sum of fragments (kb)	Identity * of fragment (site1-site2)
***Bam* H1 + *Dra* I** *Dra I in vector at:* *1910, 1929, 2621* *BamH1 at 689*	7	1 + 2	3 + 0	3.80 3.40 2.60 1.00 0.85 0.72 0.70	13.1	B - B B3 - VD1 B - B VD2 - Po B - B VD1 - VD2 VD1 - VD2
***Cla* I + *Dra* I** *Dra I in vector at:* *1910, 1929, 2621* *Cla I at 725*	7	1 + 2	4 + 0	4.40 2.50 1.90 1.50 1.25 1.00 0.70	13.3	C2 - Vo C - C C - C C - C D2 - C1 Co - D1 D1 - D2
***Pst* I + *Dra* I** *Dra I in vector at:* *1910, 1929, 2621* *Pst I at 701*	6	1 + 2	3 + 0	4.80 2.60 2.40 1.10 1.00 0.72	12.6	Po - P1 P1,2-P2,3 P1,2-P2,3 P3 - VD1 VD2 - Po VD1 - VD2
***Pst* I + *Xho* I**	6	1 + 1	3 + 1	4.75 3.10 2.40 1.10 0.80 0.72	Vector ex 2.6 ex 2.6 12.9	Po - P1 P3 - Po P1 - P2 Xh1 - P2
***Pst* I + *Hin*d III**	5!	1 + 1	3 + 1	4.20 3.10 2.60 2.40 0.70	13.0	H - P1 P3 - Vo P2 - P3 P1 - P2 Vo - H
***Xho* I + *Dra* I** *Dra I in vector at:* *1910, 1929, 2621* *Xho I at 740*	4	1 + 3	1 + 0	8.00 2.20 1.00 0.72	11.9	Xo - X1 X1 - VD1 VD2 - Xo VD1 - VD2
***Pst* I+ *Xba* I**	6	1 + 1	3 + 3	3.10 2.40 1.88 1.80 1.75 1.10	12.0	Xb3 - Vo P2 - P3 Xb2 - P2 Xb - Xb P3 - Xb3 Xb - Xb
***Xho* I + *Xba* I**	4	1 + 1	1 + 3	5.30 3.50 1.75 1.10 0.70	12.4	Xba2-Xho Xba3-Vo Xba-Xba Vo - Xba1 Xho-Xba3
***Sph* I+ *Eco* R1**	5	0 + 1	5 + 0	3.40 2.40 1.80 1.25 1.20	10.1	S5 - Vo Eo - S1

Average sum of fragments:	12.35	kb
***less* Size of vector:**	2.95	kb
Calculated size of cloned insert:	**9.40**	**kb**

* Cleavage sites are designated by the first letter of the relevant enzyme, followed by the order of the site(s) in the cloned insert, e.g. P_3 - P_0 represents the fragment spanning *Pst* I site (3) to the 3' end of the insert, which was cleaved at Pst I site (0) within the multiple cloning site of the vector. Similarly, P3 - X3 represents the fragment spanning Pst I site (3) and *Xba* I site (3).

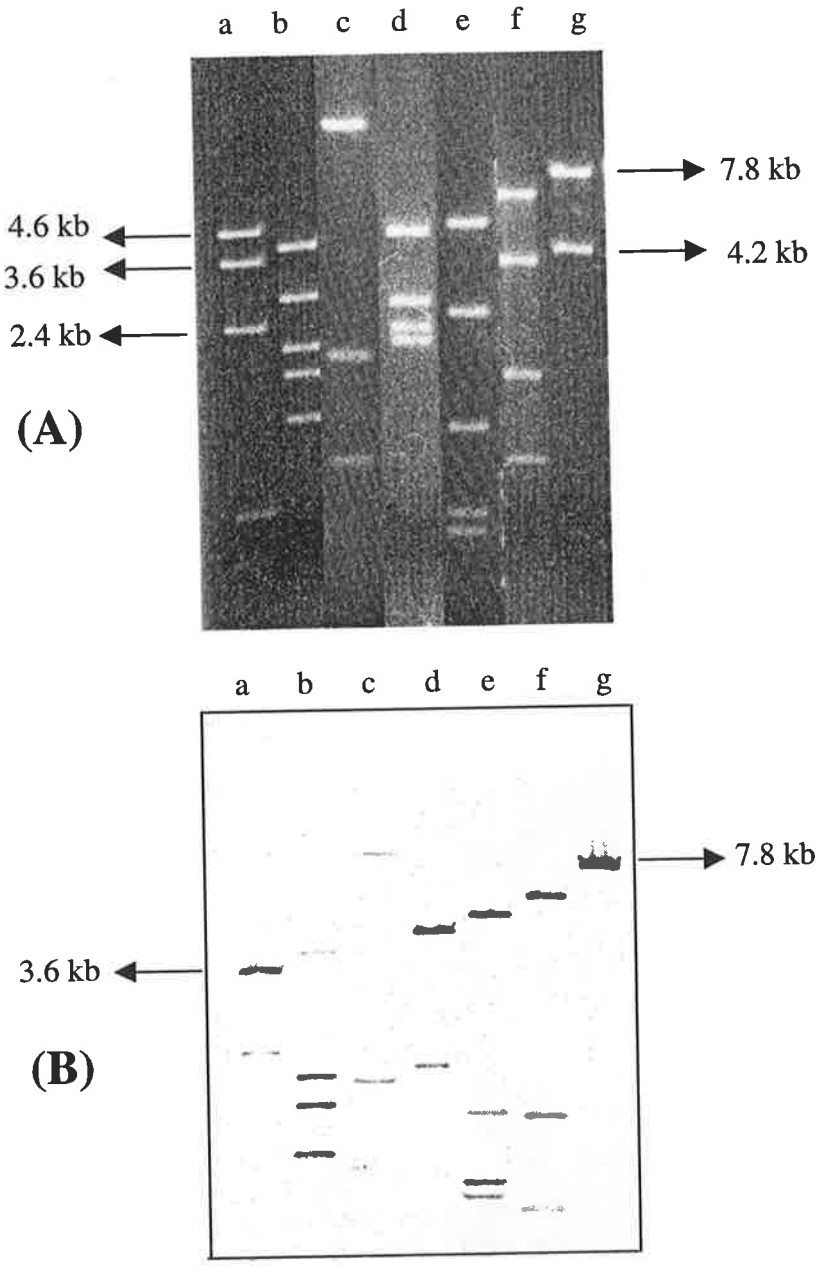


Figure 4.9. Single-enzyme restriction and Southern hybridisation analysis of pM7-1 clone.
Replicate aliquots of plasmid DNA were incubated with different restriction endonucleases. The reaction mixtures were subjected to electrophoresis on 1% agarose and stained with ethidium bromide (A). The gel was subsequently analysed by Southern hybridisation using the DIG-labelled M165-[3'] probe (B). Size markers are indicated. The enzymes used were :

a	*Bam* H I	e	*Sph* I
b	*Cla* I	f	*Xba* I
c	*Kpn* I	g	*Xho* I
d	*Pst* I		

fully characterise those segments of the 9.5-kb genomic insert that were known from hybridisation analyses to be related to the M165 probe.

The pM7-1 construct (Construct 'C') contained an insert of approximately 9.5 kb, of which 6,162 bp of contiguous sequence (from the 3' end of the insert) was compiled (**Fig. 4.11**). This sequence was found to contain four ORF, designated Sac-C/1 ('CF1'; > 1,671 bp; incompletely sequenced), Sac-C/2 ('CF2'; 1,821-bp), Sac-C/3 ('CF3'; 1,770 bp), and Sac-C/4 ('CF4'; >761-bp). The latter (CF4) was truncated by the 3' *Sac* I cloning site (**Figs. 4.10 and 4.11**). Comparative analysis of the inferred polypeptides indicated that all four ORF encoded VSP with characteristic 'CXXC' motifs and (except for the Sac/C-4 ORF 'CF4', which was truncated by the *Sac* I cloning site, an intact VSP C-terminal segment. Each ORF possessed a normal vsp gene 3' end, terminating with a single stop codon followed by the vsp gene-specific extended polyadenylation signal sequence, 'CTTAGRTAGTRAA' (Svärd *et al.* 1998; Ey *et al.* 1999). However, the CF2 (Sac/C-2), CF3 (Sac/C-3) and CF4 (Sac/C4) ORF's lacked functional 5' ends, i.e. they did not encode an N-terminal hydrophobic leader sequence that is a basic feature of all characterised VSP in *Giardia*. They therefore appeared to be pseudo genes. This was an unexpected finding, since only one definite vsp pseudo gene sequence has been reported in the published literature (Upcroft *et al.* 1993b) and this was highly corrupted, containing multiple internal stop codons. In contrast, none of the four ORF identified in the pM7-1 insert contained an internal stop codon and all appeared to be uncorrupted by other mutations, except for the absence of the immediate 5' end of what could otherwise be considered to be intact functional genes (**Fig. 4.11**). Similarity searches of the Sac-C/2, Sac-C/3 and Sac-C/4 coding sequences (and their inferred polypeptides) using FASTA and BLAST against all other available vsp gene sequences (or the encoded amino acid sequences) showed that all three are paralogues of *vsp*1269 (*crp*72), a gene described by Adam *et al.* (1992). In contrast, the incomplete 1,671-bp Sac-C/1 coding sequence showed

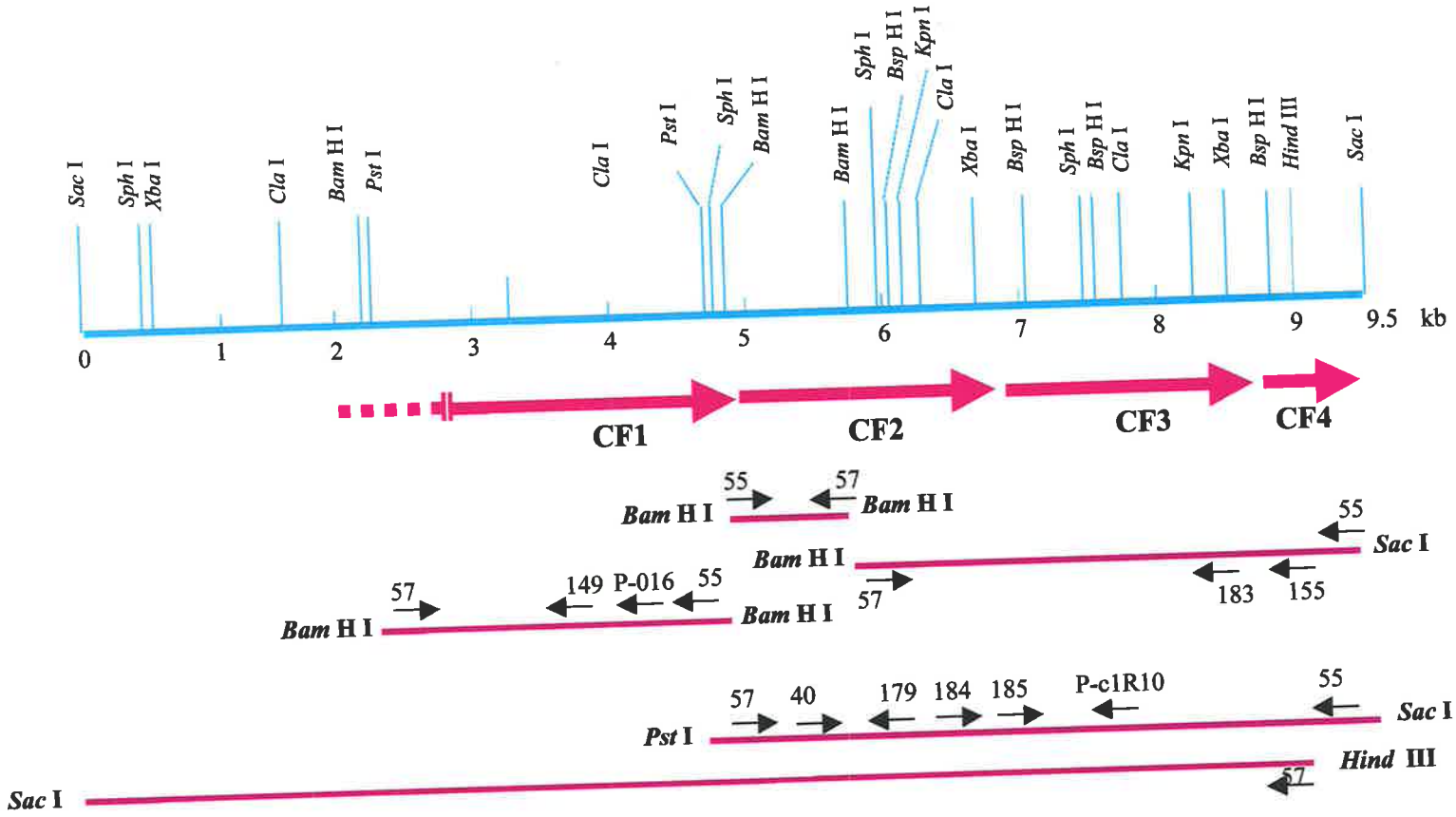


Figure 4.10. Schematic representation of the pM7-1 insert. Open reading frames (CF1, CF2, CF3, CF4) are denoted by thick arrows. The dotted line represents part of the CF1 open reading which was not fully sequenced. Subcloned fragments (used for sequencing) are depicted by bold lines. Restriction sites that were identified experimentally by cleavage are shown. The identity and annealing position of primers (used for sequencing) are shown by short arrowheads.

Figure 4.11. Nucleotide and deduced amino acid sequences specified by open reading frames identified within the 6,162-bp insert of pM7-1 (construct C). The four open reading frames CF1, CF2, CF3 and CF4 are indicated, together with the encoded C-terminal invariant segment 'CRGKA', and stop codons (both underlined), polyadenylation signal sequence (AGTRAA, double underlined). The VSP-specific pre-polyA signal sequence (CTTAGRT) is also underlined. In-frame stop codons that precede the open reading frames are highlighted. This sequence was submitted to the GeneBank™ database with accession number AF236019.

CF1

```
   1  C  T  A  C  G  D  S  K  I  V  K  T  A  K  D  Q  A  T  S  C  V  T  E
   1  TGTACTGCGTGCGGCGATAGCAAGATCGTCAAGACGGCCAAGGACCAAGCCACTTCCTGCGTGACAGAA
  24  K  E  C  T  D  A  P  G  F  F  V  D  T  T  S  G  K  K  C  S  K  C  A
  70  AAAGAGTGCACGGATGCCCCAGGCTTCTTCGTTGATACGACGAGCGGTAAAAAATGCAGCAAGTGCGCT
  47  E  T  C  K  T  C  K  T  E  A  A  K  C  T  S  C  K  D  D  K  P  Y  L
 139  GAGACATGCAAGACCTGTAAGACCGAGGCTGCGAAGTGCACCTCCTGTAAGGACGACAAGCCGTACCTG
  70  K  K  D  D  E  S  T  T  G  T  C  V  D  A  N  G  C  P  A  T  H  Y  V
 208  AAGAAGGATGACGAGTCCACAACTGGCACGTGCGTTGACGCAAACGGCTGCCCGGCAACTCATTACGTT
  93  D  K  E  A  K  E  C  N  T  C  V  S  G  G  T  V  D  C  T  T  C  E  T
 277  GATAAGGAAGCAAAGGAGTGCAACACCTGCGTTTCAGGAGGGACGGTGGACTGCACCACTTGTGAGACA
 116  S  A  N  G  V  X  C  K  X  C  X  I  X  X  K  T  K  F  A  L  G  K  N
 346  AGCGCCAACGGAGTGNTTTGTAAGANGTGTANGATTGNTANGAAGACCAAGTTTGCCCTTGGAAAGAAC
 139  S  C  V  E  N  C  P  S  N  S  S  D  E  K  T  L  G  T  C  E  C  V  E
 415  TCATGCGTTGAGAACTGCCCTTCGAACTCCAGCGATGAAAAGACCCTTGGCACCTGTGAGTGCGTCGAG
 162  G  Y  V  P  D  D  A  G  T  G  C  T  K  K  S  D  P  Q  C  N  T  P  G
 484  GGTTACGTTCCGGACGACGCGGGCACCGGGTGTACGAAGAAATCCGACCCCCAGTGCAACACCCCCGGC
 185  C  K  T  C  S  D  P  K  K  D  K  E  V  C  T  E  C  E  D  P  K  A  L
 553  TGCAAGACGTGCAGTGATCCGAAGAAAGACAAGGAGGTGTGCACAGAATGCGAAGACCCCAAGGCCCTC
 208  T  P  T  G  Q  C  I  D  N  C  G  D  L  G  G  Y  Y  E  G  T  N  E  G
 622  ACGCCCACGGGCCAGTGCATCGATAACTGCGGAGATCTGGGAGGCTACTACGAGGGCACCAACGAGGGG
 231  G  K  K  A  C  K  K  C  E  V  E  N  C  F  L  C  N  G  Q  G  Q  C  E
 691  GGCAAGAAGGCCTGCAAGAAGTGCGAGGTCGAGAACTGCTTCCTGTGCAATGGGCAGGGACAGTGCGAG
 254  T  C  K  D  G  Y  Y  K  S  G  A  A  C  A  K  C  D  T  S  C  K  T  C
 760  ACCTGCAAGGACGGGTACTACAAGAGCGGAGCCGCCTGTGCCAAGTGCGATACTTCGTGCAAGACGTGC
 277  A  N  G  K  P  N  G  C  T  S  C  E  P  K  Q  V  L  S  Y  E  G  E  G
 829  GCGAACGGGAAGCCCAACGGGTGTACGAGCTGCGAGCCTAAGCAGGTCCTCAGCTACGAAGGGGAGGGC
 300  T  G  T  Y  K  P  E  C  K  P  V  S  G  G  K  D  G  T  C  K  S  C  D
 898  ACGGGGACGTACAAACCAGAATGCAAGCCAGTGAGCGGCGGCAAGGATGGAACGTGCAAGAGCTGCGAC
 323  L  S  I  D  G  T  S  Y  C  S  A  C  N  V  G  T  E  Y  P  E  N  G  V
 967  CTGAGCATAGACGGGACAAGCTACTGTTCTGCCTGTAACGTGGGCACGGAGTATCCAGAGAACGGTGTG
 346  C  V  K  K  S  A  R  T  A  S  C  Q  V  E  P  S  N  G  V  C  G  T  C
1036  TGCGTCAAGAAGTCGGCCCGCACAGCTTCTTGCCAGGTCGAACCGAGCAATGGTGTGTGCGGGACATGT
 369  A  K  G  F  F  R  M  N  G  G  C  Y  E  T  T  K  L  P  G  K  N  V  C
1105  GCAAAGGGCTTCTTCCGCATGAACGGGGGCTGCTACGAAACGACCAAACTCCCTGGAAAGAACGTCTGT
 392  E  E  A  A  P  T  G  D  T  C  Q  T  P  A  D  G  Y  K  L  N  N  G  A
1174  GAGGAGGCAGCACCGACCGGCGATACCTGTCAGACTCCGGCCGACGGGTACAAACTGAATAACGGCGCG
 415  L  I  T  C  S  A  G  C  K  T  C  T  S  Q  D  Q  C  D  T  C  K  A  G
1243  CTCATCACTTGCTCGGCCGGATGTAAGACGTGCACCAGCCAGGACCAGTGCGACACGTGTAAGGCTGGA
 438  Y  A  K  T  G  G  N  T  K  K  C  V  P  C  A  T  G  C  S  E  C  N  A
1312  TATGCTAAGACTGGCGGTAACACTAAGAAGTGCGTTCCCTGCGCCACTGGGTGCTCCGAGTGCAATGCG
 461  D  D  A  T  K  C  T  V  C  A  A  G  Y  Y  L  S  K  E  K  C  I  A  C
1381  GACGACGCCACCAAGTGCACGGTGTGCGCTGCAGGGTACTACCTGTCCAAAGAAAAGTGCATAGCATGC
 484  D  K  S  D  G  G  S  I  T  G  V  A  N  C  A  N  C  A  P  P  T  N  N
1450  GACAAGAGCGACGGCGGATCCATCACCGGCGTCGCCAACTGCGCCAACTGCGCTCCCCCAACCAACAAT
 507  K  G  P  V  L  C  Y  L  I  Q  N  T  N  R  S  G  L  S  T  G  A  I  A
1519  AAAGGGCCTGTCCTCTGCTACCTCATACAGAACACCAACAGGAGCGGGCTTTCCACGGGGGCCATAGCG
 530 G  I  S  V  A  V  I  A  V  V  G  G  L  V  G  F  L  C  W  W  F  I  C
1588  GGGATCTCCGTCGCTGTCATCGCTGTCGTGGGGGGCCTCGTGGGCTTCCTCTGCTGGTGGTTCATATGT
```

CF2

```
      R  G  K  A  *        *  *  M  L  Y  S  D  S  D  S  D  S  A  M  W  C
1657  AGAGGGAAGGCGTGACTTAGGTAGTGAATGCTGTACAGTGATAGTGATAGTGATAGTGCTATGTGGTGT
  15  G  G  D  N  G  G  V  G  L  F  F  M  G  G  C  Y  K  T  E  S  Q  P  G
```

```
1726  GGGGGTGATAATGGTGGTGTGGGACTTTTCTTCATGGGCGGGTGCTACAAGACAGAGAGTCAGCCCGGC
  38  S  D  I  C  T  A  A  S  N  G  V  C  T  T  C  K  A  D  N  G  L  F  K
1795  AGTGACATATGCACAGCAGCCAGTAATGGAGTATGCACTACTTGTAAGGCAGATAACGGCCTGTTCAAG
  61  N  P  V  T  L  Q  K  P  G  S  E  C  I  L  C  S  D  T  K  G  A  D  G
1864  AACCCAGTTACACTTCAGAAGCCTGGTAGCGAGTGCATCCTGTGCTCTGATACTAAGGGAGCAGATGGA
  84  Y  T  G  V  A  N  C  L  K  C  T  E  P  T  N  S  P  G  A  A  T  C  T
1933  TATACGGGGGTGGCCAACTGCCTCAAGTGCACAGAACCAACTAATAGTCCAGGAGCTGCCACGTGCACT
 107  E  C  Q  E  G  Y  Y  K  E  N  N  E  C  N  Q  C  D  Q  S  C  L  T  C
2002  GAGTGTCAAGAGGGGTACTATAAAGAGAATAACGAATGCAATCAGTGTGATCAGTCCTGTCTGACTTGC
 130  S  G  S  G  P  N  H  C  T  S  C  K  E  G  K  Y  L  K  S  D  N  T  C
2071  AGTGGAAGTGGTCCGAATCATTGCACATCCTGTAAGGAGGGGAAGTATCTGAAAAGCGATAATACCTGT
 153  S  P  T  C  E  G  N  T  Y  A  D  P  V  T  R  T  C  K  E  S  G  I  A
2140  TCACCTACATGTGAAGGAAACACTTATGCTGACCCTGTGACAAGGACATGCAAGGAGTCTGGTATAGCA
 176  D  C  T  `  C  E  Y  N  T  T  V  S  K  P  Q  C  T  A  C  N  N  K  K
2209  GATTGCACCGCGTGTGAGTACAACACGACTGTCAGCAAGCCCCAGTGCACCGCCTGTAACAATAAGAAG
 199  V  K  T  E  L  D  G  T  T  T  C  V  D  D  A  G  C  A  T  D  N  V  D
2278  GTGAAGACAGAGCTGGATGGGACAACCACTTGCGTCGATGATGCTGGGTGCGCAACAGATAACGTCGAC
 222  G  S  H  F  L  S  D  D  S  T  K  C  I  L  C  S  D  T  T  T  G  T  E
2347  GGATCCCACTTTCTGAGTGATGATAGCACCAAGTGCATATTATGCAGTGATACAACAACAGGAACAGAA
 245  A  N  D  K  G  I  A  N  C  K  T  C  K  K  N  G  A  K  P  T  C  S  A
2416  GCAAACGACAAGGGGATTGCTAATTGTAAAACGTGTAAGAAGAACGGAGCCAAGCCGACCTGCTCAGCT
 268  C  L  D  G  Y  F  G  S  D  T  C  T  A  C  G  A  N  C  A  T  C  S  A
2485  TGCCTCGACGGATACTTTGGTTCTGATACATGCACAGCTTGCGGTGCCAACTGTGCTACATGCTCTGCT
 291  A  G  N  D  Q  C  T  K  C  K  P  G  F  F  M  K  Q  P  G  S  T  G  E
2554  GCTGGCAATGATCAGTGCACTAAATGCAAGCCTGGATTCTTCATGAAACAACCCGGCTCTACTGGTGAG
 314  C  V  A  C  D  S  K  A  D  N  G  I  E  G  C  S  A  C  T  N  D  G  G
2623  TGTGTTGCTTÑCGACAGTAAAGCCGACAATGGCATCGAGGGATGCAGTGCATGCACAAACGATGGCGGT
 337  A  F  K  C  T  D  C  K  P  N  Y  R  K  E  G  S  T  G  S  V  T  C  T
2692  GCTTTCAAGTGCACCGATTGCAAGCCCAACTATCGCAAGGAGGGCAGTACTGGCTCTGTCACATGCACC
 360  K  T  C  E  D  D  T  T  C  G  G  T  S  G  A  C  D  A  M  I  I  D  D
2761  AAGACCTGTGAAGATGATACTACCTGTGGTGGTACCTCTGGCGCCTGTGATGCCATGATAATAGATGAT
 383  Q  G  T  T  K  H  Y  C  S  Y  C  G  D  S  S  Q  A  P  I  D  G  L  C
2830  CAGGGCACCACAAAGCACTATTGCTCATACTGTGGGGATAGTTCACAAGCTCCTATCGATGGTCTCTGT
 406  A  T  D  K  N  G  N  T  C  T  N  N  V  C  T  S  C  T  Q  G  Y  F  L
2899  GCTACTGATAAGAATGGGAATACTTGTACCAATAATGTCTGTACATCATGCACTCAAGGGTACTTCTTG
 429  Y  M  D  G  C  Y  S  T  Q  S  Q  P  G  N  L  M  C  K  T  A  N  N  G
2968  TACATGGACGGCTGTTATAGTACTCAGAGTCAACCTGGTAACCTCATGTGCAAGACAGCTAACAACGGG
 452  V  C  T  A  V  N  E  N  N  K  Y  F  I  V  P  E  A  K  P  T  Q  Q  S
3037  GTCTGTACAGCCGTTAATGAGAATAATAAGTACTTCATAGTACCAGAAGCAAAGCCTACTCAGCAGTCT
 475  V  L  A  C  G  N  P  L  G  T  L  V  D  T  K  A  Y  V  G  V  D  G  C
3106  GTACTAGCATGTGGTAACCCATTAGGCACACTAGTAGATACTAAGGCATATGTGGGGGTGGATGGCTGC
 498  S  Q  C  T  A  P  T  A  P  S  E  G  G  M  T  P  A  V  C  T  S  C  D
3175  TCCCAGTGTACAGCCCCAACAGCTCCATCTGAGGGTGGCATGACCCCCGCTGTGTGTACCTCCTGTGAC
 521  S  G  K  K  P  N  R  D  G  T  G  C  V  T  C  S  D  T  N  C  K  S  C
3244  AGTGGTAAGAAGCCCAACAGGGATGGCACTGGGTGTGTGACATGCTCTGACACCAACTGCAAGAGCTGT
 544  A  M  D  G  V  C  G  E  C  N  S  G  F  S  L  D  N  G  K  C  V  S  T
3313  GCGATGGACGGTGTGTGTGGGGAATGCAACAGTGGCTTCAGTCTAGACAATGGCAAGTGCGTGTCCACT
 567  G  A  N  R  S  G  L  S  T  G  A  I  A  G  I  S  V  A  V  V  V  V  V
3382  GGCGCCAACAGGAGCGGGCTCTCCACGGGGGCCATCGCGGGCATCTCCGTCGCCGTGGTCGTCGTCGTG
 590  G  G  L  V  G  F  L  C  W  W  F  V  C  R  G  K  A  *
3451  GGGGGCCTCGTGGGCTTCCTCTGCTGGTGGTTCGTGTGCAGGGGGAAGGCGTGACTTGGGTAGTGAATG
CF3                                                     *  *  W  S  E  I  C  T  Q
3520  CTGTACAGTGGGTGATAATGGGTGATAGAAGGTGATAGAAGGTGATAGTGGAGTGAGATATGCACACAG
   8  A  S  D  G  L  C  T  A  C  K  A  A  N  Q  Y  I  F  Q  N  R  A  E  R
3589  GCTAGTGATGGACTGTGCACTGCTTGTAAGGCAGCTAACCAGTACATATTCCAGAACAGGGCAGAGAGG
  31  V  T  P  G  G  E  C  I  L  C  H  D  A  T  G  A  N  E  N  K  G  V  A
3658  GTGACCCCTGGTGGGGAGTGCATCCTCTGTCATGATGCCACGGGAGCGAATGAGAACAAAGGAGTGGCT
  54  N  C  L  K  C  T  A  P  A  N  S  P  G  A  A  T  C  T  E  C  M  S  G
3727  AACTGCCTCAAGTGCACGGCACCAGCTAATAGTCCAGGGGCTGCCACGTGTACTGAGTGCATGTCTGGC
  77  F  S  G  D  S  C  A  T  Q  C  G  G  N  C  A  A  C  D  K  S  N  Q  N
3796  TTCTCAGGAGACTCGTGTGCTACTCAATGCGGAGGAAACTGTGCCGCCTGTGATAAGAGTAACCAAAAT
 100  Q  C  T  S  C  K  T  G  K  Y  L  K  E  N  Q  C  V  E  K  S  A  C  N
3865  CAGTGCACCTCTTGTAAGACTGGGAAGTACCTGAAAGAGAATCAATGTGTGGAAAAAAGTGCATGTAAC
 123  N  N  H  Y  P  D  D  T  S  M  T  C  V  A  C  S  T  I  Q  D  C  T  A
```

```
3934  AACAATCACTACCCCGATGACACTAGCATGACCTGCGTCGCCTGTAGCACCATACAAGATTGCACAGCA
 146  C  T  M  D  Q  S  T  G  K  P  K  C  T  N  C  G  A  S  K  I  P  R  T
4003  TGCACAATGGACCAGTCCACGGGGAAGCCAAAGTGCACTAACTGCGGTGCCAGTAAGATCCCCAGGACG
 169  T  L  D  G  T  S  T  C  V  T  K  G  Y  D  Q  C  Q  G  T  D  K  E  L
4072  ACTCTCGACGGAACCTCGACCTGCGTGACAAAAGGGTATGATCAGTGTCAAGGCACAGATAAGGAGCTG
 192  F  M  K  E  D  Q  S  A  C  L  L  C  G  D  N  T  E  D  T  S  E  S  N
4141  TTCATGAAGGAGGACCAATCGGCGTGCTTATTATGCGGAGACAATACTGAGGATACATCTGAGTCCAAT
 215  K  N  Q  G  T  P  N  C  K  T  C  T  K  T  A  S  T  K  P  V  C  E  T
4210  AAAAACCAGGGAACTCCTAATTGCAAGACCTGTACCAAGACAGCATCCACTAAACCGGTCTGTGAGACG
 238  C  K  D  G  F  F  F  N  T  G  D  K  T  C  T  N  K  C  D  A  T  C  K
4279  TGCAAGGATGGCTTCTTCTTTAATACTGGTGATAAGACATGTACAAACAAGTGCGATGCCACCTGCAAG
 261  T  C  S  A  A  T  D  A  N  R  C  L  T  C  M  P  G  Y  F  P  I  D  S
4348  ACGTGCTCTGCTGCTACAGATGCTAACCGATGCTTAACRTGTATGCCCGGATACTTCCCCATCGATAGT
 284  T  D  Q  Q  G  K  K  C  V  P  C  D  S  V  D  D  K  G  R  E  G  C  S
4417  ACAGATCAACAAGGGAAGAAGTGCGTCCCATGCGATAGTGTGGACGACAAGGGTCGCGAAGGGTGCAGT
 307  V  C  S  N  N  G  G  F  K  C  T  E  C  K  V  N  Y  K  K  Q  S  N  G
4486  GTCTGCTCTAACAATGGTGGATTCAAGTGTACTGAATGTAAGGTGAACTATAAGAAGCAGTCCAATGGA
 330  D  A  G  D  D  Y  T  C  V  K  T  C  E  D  P  T  A  C  G  G  T  S  G
4555  GATGCAGGGGATGACTATACATGTGTGAAGACTTGTGAAGACCCCACGGCTTGTGGTGGGACATCTGGC
 353  A  C  D  A  I  V  I  D  N  D  G  V  E  H  H  Y  C  S  Y  C  G  N  Q
4624  GCCTGCGATGCTATAGTAATCGACAATGATGGAGTAGAGCATCATTACTGTTCTTATTGTGGAAATCAA
 376  G  D  V  P  I  N  G  I  C  T  S  S  L  A  S  N  T  C  A  D  G  V  C
4693  GGAGATGTACCAATCAATGGTATATGCACCAGCTCACTTGCAAGTAATACCTGCGCCGATGGTGTCTGC
 399  K  S  C  A  Q  G  Y  F  M  Y  M  G  G  C  Y  D  V  S  K  A  P  G  S
4762  AAGTCTTGTGCTCAAGGGTACTTTATGTATATGGGAGGGTGCTATGATGTGTCGAAGGCTCCTGGTAGT
 422  H  M  C  K  K  A  D  N  N  G  I  C  T  E  A  A  S  N  R  Y  F  V  V
4831  CACATGTGCAAGAAAGCAGATAACAATGGGATATGCACAGAGGCGGCAAGTAATAGGTACTTCGTAGTC
 445  P  G  A  Q  T  T  D  Q  S  V  L  A  C  G  N  P  L  G  T  L  V  G  T
4900  CCTGGGGCACAGACCACTGATCAGTCTGTATTAGCCTGTGGCAACCCCCTAGGCACACTAGTAGGCACT
 468  Q  G  T  A  K  A  Y  V  G  V  D  G  C  S  Q  C  T  A  P  A  G  L  S
4969  CAAGGTACCGCTAAGGCATACGTGGGGGTGGATGGCTGCTCACAGTGCACAGCCCCAGCAGGTCTGTCT
 491  E  G  G  M  T  P  A  V  C  T  S  C  D  N  S  K  K  P  N  K  D  G  S
5038  GAGGGTGGCATGACCCCCGCTGTCTGTACCTCCTGTGACAATAGTAAGAAGCCCAACAAGGATGGCAGC
 514  G  C  V  L  C  S  D  T  N  C  K  S  C  V  M  D  G  V  C  G  E  C  N
5107  GGGTGTGTGCTGTGCTCTGACACCAACTGCAAGAGCTGTGTGATGGACGGTGTGTGTGGGGAGTGCAAC
 537  S  G  F  S  L  D  N  G  K  C  V  S  S  G  A  N  R  S  G  L  S  T  G
5176  AGTGGCTTCAGTCTAGACAATGGTAAGTGTGTGTCCTCTGGCGCCAACAGGAGCGGGCTCAGCACGGGC
 560  A  I  A  G  I  S  V  A  V  I  A  V  V  G  G  L  V  G  F  L  C  W  W
5245  GCCATCGCGGGGATCTCCGTGGCTGTCATCGCTGTCGTGGGGGGCCTCGTGGGCTTCCTCTGCTGGTGG
 583  F  V  C  R  G  K  A  *
5314  TTCGTGTGCAGGGGGAAGGCGTGACTTGGGTAGTGAATGCTGTACAGTGGGTGATAATGGGTGATAGAA
CF4                    *  *  W  S  V  I  C  T  A  A  S  D  G  V  C  T  A  C  N
5383  GGTGATAGAGGGTGATAATGGAGTGTGATATGTACAGCAGCCAGTGATGGAGTGTGCACTGCTTGTAAC
  18  T  A  N  Q  Y  I  F  Q  N  K  A  T  T  V  T  P  G  G  E  C  I  L  C
5452  ACAGCTAACCAGTACATATTCCAGAACAAGGCTACGACAGTGACCCCTGGTGGGGAGTGCATCCTCTGT
  41  H  D  A  T  G  A  N  E  N  K  G  V  A  N  C  L  K  C  T  E  P  S  G
5521  CATGATGCCACGGGAGCGAATGAGAACAAAGGAGTGGCTAACTGCCTCAAGTGCACGGAACCAAGTGGT
  64  S  P  G  A  A  T  C  T  E  C  Q  A  G  Y  Y  K  D  G  N  G  A  C  V
5590  AGTCCAGGAGCTGCCACGTGTACAGAGTGTCAAGCAGGGTACTATAAAGATGGTAACGGTGCATGTGTC
  87  K  C  N  E  A  C  L  T  C  S  G  E  G  A  T  K  C  T  F  C  A  G  E
5659  AAATGTAATGAAGCTTGTCTGACTTGCAGCGGTGAGGGCGCAACTAAGTGCACCTTCTGTGCGGGCGAG
 110  K  Y  L  D  G  N  N  C  V  D  A  S  S  C  N  G  D  K  Y  P  N  P  S
5728  AAGTATCTCGATGGGAATAATTGTGTAGATGCTAGCAGTTGTAATGGAGATAAGTACCCCAACCCAAGC
 133  T  G  K  C  T  A  C  N  A  G  A  D  Q  G  G  I  P  E  C  T  A  C  T
5797  ACTGGGAAGTGCACAGCTTGTAACGCTGGAGCAGATCAGGGGGGCATACCAGAGTGCACTGCTTGCACG
 156  Y  N  A  K  L  Q  K  P  V  C  S  A  C  N  G  K  K  V  K  T  E  L  D
5866  TACAATGCAAAGCTACAGAAGCCGGTGTGCAGTGCCTGTAACGGTAAGAAGGTGAAGACAGAGCTGGAT
 179  G  T  T  T  C  V  D  D  A  G  C  K  K  D  S  T  H  F  V  D  D  D  N
5935  GGGACAACAACCTGTGTCGATGATGCTGGGTGCAAAAAAGACAGCACACACTTTGTTGACGATGATAAT
 202  T  M  C  V  L  C  S  D  T  S  G  N  E  P  N  K  N  K  G  L  K  G  C
6004  ACTATGTGTGTTCTGTGCAGCGATACCAGCGGAAACGAGCCAAATAAGAACAAGGGACTCAAAGGCTGT
 220  K  A  C  T  K  Q  S  S  S  P  P  T  C  T  G  C  L  P  G  Y  Y  D  S
6073  AAGGCATGTACTAAGCAGTCCAGTAGTCCCCCCACGTGTACCGGATGCGGCTCGTTTGTTTGTGAGCTC
                                                                    Sac I
```

little similarity to *crp*72 (only 49% nt identity over 1,678 bp) or to the other vsp ORF identified in the pM7-1 insert.

4.8.2b Construct A (pM11-2)

The genomic inserts recovered in pM3-1D and pM11-2 were also similar in size. Each was characterised by restriction mapping (as exemplified in **Fig. 4.12** and **Table 4.5**), Southern hybridisations (data summarised for pM3-1D in **Table 4.5**) and partial sequence determinations. The accumulated data (not shown) indicated that the two inserts were identical, but cloned in opposite directions. Consequently, pM11-2 was chosen for further analysis and characterisation. Use of the pM11-2 plasmid as a template for primer-based sequencing reactions proved unsatisfactory, presumably because of false priming within multiple related vsp gene sequences whose presence had been indicated by Southern hybridisation data (data not shown). In order to circumvent this problem, nine subclones with inserts ranging from 0.5 to 3.2 kb were constructed from pM11-2. These are depicted in **Fig. 4.13**. Each subclone was initially sequenced from both ends of the respective insert using T3 and T7 primers. Additional primers were designed from the data and used to further sequence the inserts. Overlapping and high quality sequence data were compiled to obtain the complete 5444-bp sequence of the parent (pM11-2) insert. This is presented in **Fig. 4.14**. Four ORF, designated *Sac*-A/1 ('AF1'; >1206 bp), *Sac*-A/2 ('AF2'; 1,746 bp), *Sac*-A/3 ('AF3'; 1,746 bp) and *Sac*-A/4 ('AF4'; >547 bp) were detected in a head-to-tail arrangement (**Figs. 4.13, 4.14**). The AF1 (*Sac*-A/1) and AF4 (*Sac*-A/4) reading frames were situated at the 5' and 3' ends of the insert, respectively. Because these overlapped the *Sac* I cleavage sites, the 5' (*Sac*-A/1) or 3' (*Sac*-A/4) portion of the respective coding sequences were missing.

4.8.2c Construct D (pM13-3)

The 3.8-kb genomic fragment recovered in pM13-3 was the smallest of the five fully characterised clones. The insert was tested for cleavage by various restriction endonucleases

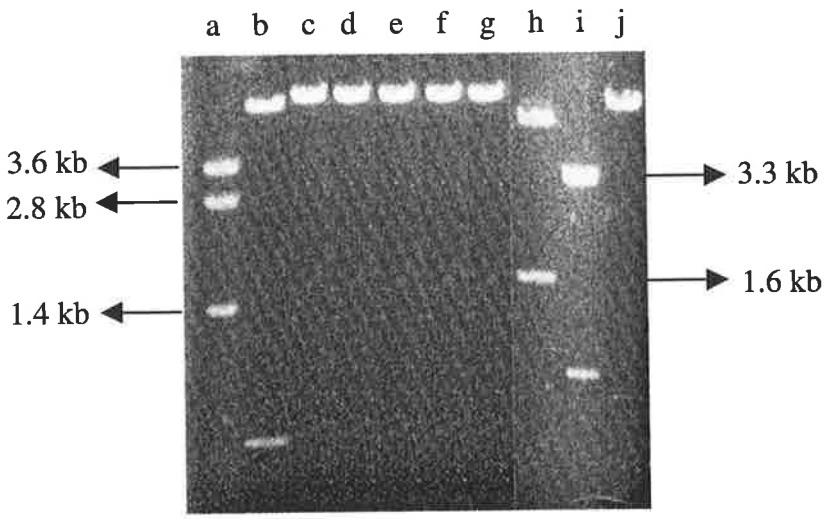


Figure 4.12. Restriction analysis of pM11-2 (construct A), containing a 5.4-kb genomic insert. In lanes c-g and j, the construct was linearised by cleavage within the multiple cloning site of the vector, i.e. the insert lacked sites for these enzymes. The endonucleases used were:

a	*Cla* I	f	*Kpn* I
b	*Dra* I	g	*Pst* I
c	*Eco* R I	h	*Sph* I
d	*Eco* R V	i	*Xba* I
e	*Hind* III	j	*Xho* I

Table 4.5. Restriction endonuclease cleavage data and associated Southern hybridisation data for plasmid construct pM3-1D (construct 'S')

A. *Single-enzyme cleavage tests*

Enzyme tested	No. of fragments	No. of sites in vector	No. of sites in insert	Fragment sizes (kb)	Sum of fragments (kb)	Identity * of fragment (Site1-Site2)	Hybridisation with probe 148-149
Eco R1	1	1	0				
Eco RV	1	1	0				
Hind III	1	1	0				
Kpn I	1	1	0				
Pst I	1	1	0				
Sca I	1	1	0				
Xho I	0	0	0				
Bam HI	2	1	1	7.0		B1 - Vo	
				0.4	7.4	Bo - B1	
Cla I	3	1	2	3.8		C1 - C2	++++
				3.0		C2 - Vo	-
				1.4	8.2	Co - C1	++++
Dra I	2	0	2	6.5		D2 - D1	
				0.7	7.2	D1 - D2	
Sph I	2	0	2	5.5		S2 - S1	++++
				1.7	7.2	S1 - S2	++++
Xba I	3	1	2	3.4			
				3.3			
				1.0	7.7		

Average sum of fragments:	7.53	kb
***less* Size of vector:**	2.95	kb
Calculated size of cloned insert:	**4.58**	**kb**

B. *Double - enzyme cleavage tests*

Enzyme tested	No. of fragments	Sites in vector	Sites in insert	Fragment sizes (kb)	Sum of fragments (kb)	Identity * of fragment (site1-site2)	Hybridisation with probe 148-149
Bam H1 + *Cla* I	3	1 + 1	1 + 2	3.50			
				2.85			
				1.50	7.9	Co - C1	
Bam H1 + *Sph* I	4	1 + 0	1 + 2	4.70		S2 - Vo	
				1.70		S1- S2	+++
				1.25		B1 - S1	
				0.40	8.1	Bo - B1	
Dra I + *Hin*d III	3	0 + 1	2 + 0	6.30		D2 - Vo	
				1.00		VoH - D1	
				0.70	8.0	D1 - D2	
Eco R1 + *Sph* I	3	1 + 0	0 + 2	4.70			
				1.70			
				1.60	8.0		
*Hin*d III + *Sph* I	3	1 + 0	0 + 2	4.70			
				1.70			
				1.60	8.0		
Pst I + *Xba* I	3	1 + 1	0 + 2	3.60			
				3.55			
				1.00	8.2		

Average sum of fragments:	8.01	kb
***less* Size of vector:**	2.95	kb
Calculated size of cloned insert:	**5.06**	**kb**

* Cleavage sites are designated by the first letter of the relevant enzyme, followed by the order of the site(s) in the cloned insert, e.g. C_0 - C_1 represents the fragment spanning *Cla* I site (1) to the 3' end of the insert, which was cleaved at ClaI site (0) within the multiple cloning site of the vector. Similarly, S_1 - S_2 represents the fragment spanning *Sma* I sites (1) and (2).

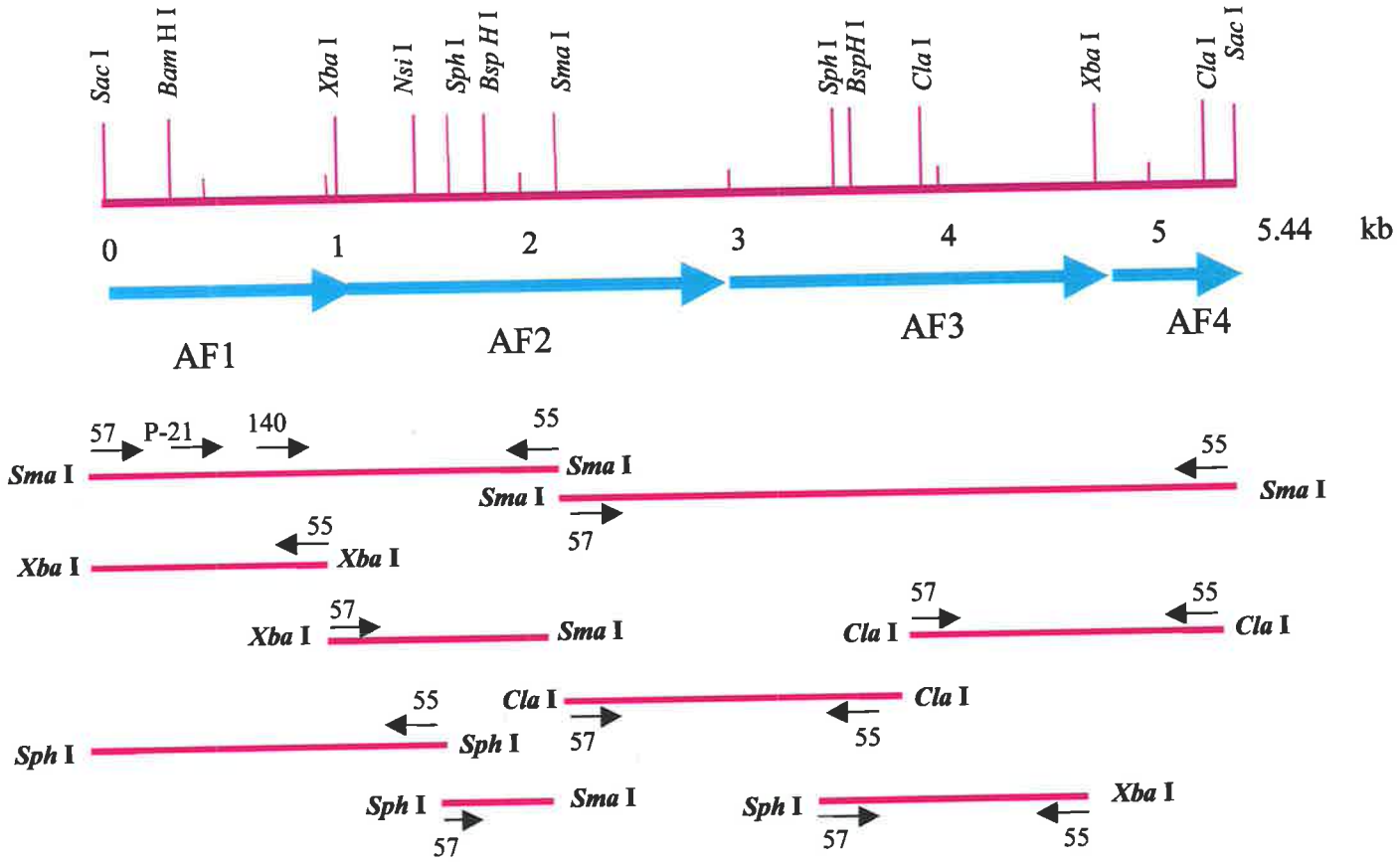


Figure 4.13. Schematic representation of the pM11-2 insert. Open reading frames AF1, AF2, AF3, AF4 are denoted by thick arrows. Subcloned fragments (used for sequencing) are depicted by bold lines. Restriction sites that were identified experimentally by cleavage and confirmed by sequence analysis, are shown. The identity and annealing positions of primers, used for sequencing and PCR, are shown by short arrowheads.

Figure 4.14. Nucleotide and deduced amino acid sequences specified by open reading frames within the 5,443-bp insert of pM11-2 (construct A). The four open reading frames AF1, AF2, AF3 and AF4 are indicated, together with encoded C-terminal invariant segment 'CRGKA' and stop codons (both underlined), polyadenylation signal sequence (AGTRAA, double underlined) and the VSP-specific pre-polyA signal sequence (CTTAGRT, underlined). In-frame stop codons that precede the open reading frames are highlighted. This sequence was submitted to the GeneBank™ database with accession number AF230275.

```
        Sac I
AF1     S  S  N  C  V  D  A  D  H  F  K  A  D  G  D  S  A  C  V  L  C  S  D
   1 GAGCTCTAATTGCGTGGATGCTGATCACTTCAAGGCAGATGGTGACTCAGCGTGTGTCTTGTGCAGCGAT
  24  T  S  G  S  D  P  K  N  K  G  I  A  G  C  K  A  C  T  K  T  A  S  A
  70 ACCAGCGGATCAGACCCAAAGAACAAAGGCATTGCTGGCTGTAAGGCTTGCACGAAGACAGCGAGTGCT
  47  N  P  T  C  S  E  C  L  A  G  Y  Y  S  S  G  S  G  T  V  T  C  T  A
 139 AATCCAACCTGCTCGGAGTGCCTAGCTGGATACTACAGTAGTGGTAGTGGTACGGTTACATGCACAGCC
  70  C  G  Q  N  C  A  T  C  T  Q  A  G  N  N  Q  C  T  K  C  K  A  G  F
 208 TGTGGACAAAACTGTGCCACCTGCACTCAAGCTGGCAACAATCAGTGCACTAAGTGCAAGGCAGGGTTC
  93  F  L  K  G  N  G  P  G  E  C  V  P  C  D  D  A  Q  N  Q  G  I  E  G
 277 TTCCTGAAAGGCAATGGTCCCGGTGAGTGTGTCCCATGTGATGATGCACAGAATCAAGGGATCGAAGGC
 116  C  A  E  C  T  G  T  T  G  S  L  K  C  T  K  C  K  P  N  R  K  P  A
 346 TGTGCCGAGTGCACTGGTACTACAGGATCCCTCAAGTGTACTAAGTGCAAGCCTAACAGGAAGCCCGCG
 139  G  N  Q  D  N  Y  T  C  T  E  K  T  C  E  D  E  T  A  C  G  G  T  A
 415 GGAAACCAAGACAACTATACATGCACAGAGAAGACCTGCGAAGATGAGACTGCATGTGGTGGGACAGCT
 162  G  S  C  D  A  T  V  I  D  N  Q  G  N  A  K  H  Y  C  S  Y  C  G  E
 484 GGCTCTTGTGATGCCACAGTGATTGATAATCAGGGCAATGCAAAGCACTACTGTTCCTATTGTGGAGAG
 185  A  N  K  F  P  I  N  G  V  C  V  E  S  S  Q  K  N  G  N  N  C  N  S
 553 GCTAATAAGTTCCCTATCAACGGTGTCTGTGTTGAAAGTAGTCAAAAGAACGGGAATAACTGTAACAGT
 208  H  T  C  T  S  C  A  Q  G  Y  F  M  Y  M  G  G  C  Y  D  A  S  K  A
 622 CATACATGTACATCATGTGCTCAAGGGTACTTTATGTACATGGGAGGGTGCTATGATGCGTCGAAGGCT
 231  P  G  N  H  M  C  K  T  A  D  N  G  V  C  T  T  P  N  A  N  N  R  Y
 691 CCCGGTAATCATATGTGCAAGACCGCTGACAACGGTGTCTGTACTACTCCCAATGCCAACAATAGGTAC
 254  F  L  V  P  G  A  S  N  T  D  Q  S  V  L  A  C  S  N  P  L  G  T  L
 760 TTCCTGGTCCCAGGGGCATCCAACACTGATCAGTCTGTGTTGGCTTGTAGTAACCCATTAGGCACACTA
 277  V  G  T  G  D  T  A  K  A  Y  V  G  V  L  N  C  S  Q  C  T  A  P  A
 829 GTAGGCACTGGTGATACTGCTAAGGCATACGTGGGGGTACTCAACTGTTCACAGTGCACAGCCCCAGCA
 300  A  L  S  D  G  G  M  A  A  A  I  C  T  S  C  D  S  S  K  V  P  N  R
 898 GCTCTGTCTGATGGTGGTATGGCTGCTGCTATATGTACCTCCTGTGACAGTAGTAAGGTGCCCAACAGG
 323  D  G  S  G  C  V  L  C  S  D  T  N  C  K  S  C  V  M  D  G  V  C  G
 967 GATGGCAGCGGGTGTGTGCTGTGCTCTGACACCAACTGCAAGAGCTGTGTGATGGACGGTGTGTGTGGG
 346  E  C  S  D  G  Y  S  L  E  G  G  R  C  V  S  S  G  A  N  R  S  G  L
1036 GAGTGCAGTGATGGGTACAGTCTAGAGGGTGGCAGGTGCGTGTCCTCTGGCGCCAACAGGAGCGGGCTC
 369  S  T  G  A  I  A  G  I  S  V  A  V  I  A  V  V  G  G  L  I  G  F  L
1105 TCCACGGGGGCCATCGCGGGGATCTCCGTCGCTGTCATCGCTGTCGTGGGTGGTCTCATAGGGTTCCTC
 392   C  W  W  F  V  C  R  G  K  A  *
1174 TGCTGGTGGTTCGTGTGTAGGGGGAAGGCGTGACTTAGGTAGTGAATGCTCTATAGTGTATGATAATGG
AF2                *  *  W  S  E  I  C  T  A  A  S  D  G  L  C  T  A  C  K
1243 GGTGATAGAGGGTGATAGTGGAGTGAGATATGCACAGCAGCCAGTGATGGACTGTGCACTGCTTGTAAG
  18  T  A  N  G  L  F  K  N  P  A  T  A  P  E  K  G  R  E  C  I  L  C  S
1312 ACAGCTAACGGCCTGTTCAAGAACCCTGCTACAGCACCAGAGAAGGGAAGGGAGTGCATCCTGTGCTCT
  41  D  T  T  D  R  D  G  V  T  G  A  A  G  C  S  E  C  S  H  T  G  T  S
1381 GATACGACAGATAGAGACGGGGTCACGGGAGCCGCTGGATGCTCAGAATGCTCCCACACGGGCACTAGC
  64  G  P  A  T  C  T  V  C  Q  D  G  Y  I  K  K  G  D  A  C  E  K  C  D
1450 GGACCGGCCACATGCACTGTCTGCCAAGATGGATACATTAAGAAGGGTGATGCATGTGAGAAGTGTGAT
  87  Q  S  C  L  T  C  D  A  S  G  P  N  Q  C  T  S  C  P  E  G  K  Y  L
1519 CAGTCTTGTCTGACCTGCGATGCAAGTGGTCCGAATCAGTGTACATCGTGTCCAGAAGGGAAGTACCTG
 110  K  S  N  Q  C  V  Q  D  T  G  C  D  G  N  T  Y  A  D  P  E  S  G  K
1588 AAGAGTAATCAATGTGTACAAGATACTGGATGCGATGGAAACACTTATGCTGACCCAGAGTCGGGGAAG
 133  C  K  A  C  S  T  I  H  A  E  C  T  A  C  S  F  D  K  G  T  G  K  P
1657 TGCAAGGCATGTAGTACGATACACGCAGAGTGCACGGCATGCTCCTTCGACAAGGGCACGGGGAAGCCA
 156  K  C  T  N  C  G  A  S  K  I  P  R  T  T  L  D  G  T  S  T  C  V  T
1726 AAGTGCACAAACTGCGGTGCCAGTAAGATCCCCAGGACGACCCTCGACGGGACCTCGACCTGCGTGACA
```

```
 179  K  G  Y  T  E  C  Q  G  A  D  K  E  L  F  M  K  E  D  Q  S  A  C  L
1795  AAAGGGTATACAGAATGCCAAGGAGCAGACAAGGAGCTGTTCATGAAGGAGGACCAATCGGCGTGCCTG
 202  L  C  G  D  T  K  E  A  S  N  D  K  G  V  A  N  C  R  T  C  T  K  N
1864  CTGTGCGGAGATACTAAAGAAGCTAGCAACGATAAGGGGGTCGCTAACTGTAGGACGTGCACTAAGAAT
 225  A  N  D  S  P  P  T  C  T  A  C  L  N  G  Y  F  F  E  Q  S  S  N  T
1933  GCAAATGACTCCCCTCCAACTTGCACAGCGTGCCTCAATGGATACTTCTTTGAACAGTCATCAAACACT
 248  C  S  A  C  G  A  N  C  A  T  C  S  A  K  T  A  E  D  K  C  L  T  C
2002  TGTTCGGCTTGTGGGGCCAACTGTGCCACATGTTCAGCTAAAACTGCTGAGGATAAGTGCTTAACATGC
 271  K  E  G  F  F  L  A  G  T  G  E  G  K  C  I  S  C  E  N  G  N  D  S
2071  AAGGAGGGGTTCTTCCTGGCCGGCACCGGCGAAGGAAAGTGCATCTCGTGCGAGAATGGTAATGATAGT
 294  G  Y  S  G  L  A  R  C  L  S  C  T  A  P  A  A  P  G  P  A  S  C  N
2140  GGCTACAGCGGGCTTGCAAGATGCCTCTCTTGCACAGCCCCAGCCGCCCCGGGCCCGGCCAGCTGCAAC
 317  R  C  K  I  G  Y  K  L  Q  G  T  T  C  V  K  T  C  E  D  E  T  A  C
2209  CGCTGCAAGATCGGCTACAAGTTACAGGGAACAACGTGTGTGAAGACCTGCGAGGATGAGACTGCCTGC
 340  G  G  T  S  G  A  C  D  A  I  V  I  D  D  K  G  N  T  K  H  Y  C  S
2278  GGGGGTACTTCTGGGGCCTGCGATGCCATCGTAATAGATGATAAGGGCAACACAAAGCACTACTGTTCA
 363  Y  C  G  D  S  S  K  Y  P  I  D  G  L  C  A  S  E  A  Q  S  N  T  C
2347  TACTGTGGGGATAGCAGTAAGTATCCTATAGATGGTCTCTGTGCTAGTGAGGCTCAAAGTAATACTTGT
 386  N  N  H  V  C  E  S  C  T  T  G  Y  F  L  Y  M  N  G  C  Y  D  V  S
2416  AACAACCATGTCTGTGAGTCTTGTACTACGGGGTACTTCCTATACATGAACGGCTGCTATGATGTGTCG
 409  K  P  P  G  N  L  M  C  S  K  A  T  T  A  G  V  C  T  E  A  A  N  N
2485  AAGCCTCCTGGTAATCTCATGTGCTCCAAAGCAACAACAGCTGGTGTCTGCACAGAAGCAGCCAACAAT
 432  K  Y  F  V  V  P  E  A  K  A  T  D  Q  S  V  L  A  C  G  N  P  L  G
2554  AAGTACTTCGTAGTCCCAGAAGCAAAGGCTACTGATCAGTCTGTACTAGCCTGTGGTAACCCCCTGGGC
 455  T  L  I  G  T  Q  G  T  A  K  A  Y  V  G  V  D  G  C  S  Q  C  T  A
2623  ACATTGATAGGCACTCAAGGTACTGCTAAGGCATACGTGGGGGTGGATGGCTGCTCCCAGTGTACAGCC
 478  P  T  Q  L  E  A  P  G  M  T  S  A  V  C  T  A  C  D  S  G  K  K  P
2692  CCGACACAGCTAGAAGCTCCTGGTATGACCTCCGCTGTGTGTACTGCCTGTGACAGTGGTAAGAAGCCC
 501  N  R  D  G  S  G  C  V  L  C  S  D  T  N  C  K  S  C  V  M  D  S  V
2761  AACAGGGATGGCAGTGGGTGTGTGCTATGCTCTGACACCAACTGCAAGAGCTGTGTGATGGACAGTGTG
 524  C  G  E  C  N  S  G  F  S  L  D  N  G  K  C  V  S  S  G  A  N  R  S
2830  TGTGGGGAGTGCAACAGTGGCTTCAGTCTAGACAATGGTAAGTGTGTGTCCTCTGGCGCCAACAGAAGC
 547  G  L  S  T  G  A  I  A  G  I  S  V  A  V  I  A  V  V  G  G  F  V  G
2899  GGGCTCTCCACGGGGGCCATCGCGGGGATCTCCGTCGCTGTCATCGCTGTCGTGGGGGGGTTCGTGGGC
 570  F  L  C  W  W  F  V  C  R  G  K  A  *
2968  TTCCTCTGCTGGTGGTTCGTGTGTAGGGGGAAGGCGTGACTTAGGTAGTGAATGCTGTACAATGTGTGAT
```

AF3

```
                                *  *  W  S  V  I  C  T  A  A  S  D  G  V  C
3037   GAAGGTGATAGTGGGTGATAGTGTGATAATGGAGTGTGATATGTACAGCAGCCAGTGATGGAGTGTGC
  14   T  A  C  N  T  A  N  Q  Y  I  F  Q  N  K  A  T  T  V  T  P  G  G  E
3106   ACTGCTTGTAACACAGCTAACCAGTACATATTCCAGAACAAGGCTACGACAGTGACCCCTGGTGGGGAG
  37   C  I  L  C  H  D  A  T  G  A  N  E  N  K  G  V  A  N  C  L  K  C  T
3175   TGCATACTGTGCCATGATGCCACGGGAGCGAATGAGAACAAGGGAGTGGCTAATTGCCTCAAGTGCACG
  60   A  P  A  N  S  P  G  A  A  T  C  T  E  C  M  S  G  F  S  G  D  S  C
3244   GCACCAGCTAATAGTCCAGGGGCTGCCACGTGTACTGAGTGCATGTCTGGCTTCTCAGGAGACTCGTGT
  83   A  T  Q  C  G  G  D  C  A  A  C  D  K  S  N  Q  N  Q  C  T  S  C  K
3313   GCTACTCAATGCGGAGGAGACTGTGCCGCCTGTGATAAGAGTAACCAAAATCAGTGCACCTCTTGTAAG
 106   T  G  K  Y  L  K  E  N  Q  C  V  E  K  S  A  C  N  N  N  H  Y  P  D
3382   ACTGGGAAGTACCTGAAAGAGAATCAATGTGTGGAAAAAAGTGCATGTAACAACAATCACTACCCCGAT
 129   D  T  S  M  T  C  V  A  C  S  T  I  Q  D  C  T  A  C  T  M  D  Q  S
3451   GACACTAGCATGACCTGCGTCGCCTGTAGCACCATACAAGATTGCACAGCATGCACAATGGACCAGTCC
 152   T  G  K  P  K  C  T  N  C  G  A  S  K  I  P  R  T  T  L  D  G  T  S
3520   ACGGGGAAGCCAAAGTGCACTAACTGCGGTGCCAGTAAGATCCCCAGGACGACTCTCGACGGAACCTCG
 175   T  C  V  T  K  G  Y  T  E  C  Q  G  A  D  K  E  L  F  M  K  E  D  Q
3589   ACCTGCGTGACAAAAGGGTATACAGAATGCCAAGGAGCAGACAAGGAGCTGTTCATGAAGGAGGACCAA
 198   S  A  C  L  L  C  G  D  N  T  E  D  T  S  E  S  N  K  N  Q  G  T  P
3658   TCGGCGTGCTTATTATGCGGAGACAATACTGAGGATACATCTGAGTCCAATAAAAACCAGGGAACTCCT
 221   N  C  K  T  C  T  K  T  A  S  T  K  P  V  C  E  T  C  K  D  G  F  F
3727   AATTGCAAGACCTGTACCAAGACAGCATCCACTAAACCGGTCTGTGAGACGTGCAAGGATGGCTTCTTC
 244   F  N  T  G  D  K  T  C  T  N  K  C  D  A  T  C  K  T  C  S  A  A  T
3796   TTTAATACTGGTGATAAGACATGTACAAACAAGTGCGATGCCACCTGCAAGACGTGCTCTGCTGCTACA
 267   D  A  N  R  C  L  T  C  M  P  G  Y  F  P  I  D  S  T  D  Q  Q  G  K
3865   GATGCTAACCGATGCTTAACTTGTATGCCCGGATACTTCCCCATCGATAGTACAGATCAACAAGGGAAG
 290   K  C  V  P  C  D  S  V  D  D  K  G  R  E  G  C  S  V  C  S  N  N  G
3934   AAGTGCGTCCCATGCGATAGTGTGGACGACAAGGGTCGCGAAGGGTGCAGTGTCTGCTCTAACAATGGT
```

```
 313  G  F  K  C  T  E  C  K  V  N  Y  K  K  Q  S  N  G  D  A  G  D  D  Y
4003  GGATTCAAGTGTACTGAATGTAAGGTGAACTATAAGAAGCAGTCCAATGGAGATGCAGGGGATGACTAT
 336  T  C  V  K  T  C  E  D  P  T  A  C  G  G  T  A  G  A  C  D  A  M  I
4072  ACATGTGTGAAGACTTGTGAAGACCCCACGGCTTGTGGGGGGACAGCTGGCGCCTGCGATGCCATGATA
 359  I  D  D  Q  G  N  T  K  H  Y  C  S  Y  C  G  E  T  N  K  I  P  I  D
4141  ATTGATGATCAGGGTAACACAAAGCACTACTGTTCCTACTGTGGAGAGACCAATAAGATACCAATTGAT
 382  G  K  C  V  D  S  G  S  I  N  G  N  T  C  N  S  H  T  C  T  S  C  A
4210  GGGAAGTGTGTTGATAGTGGTAGTATAAATGGTAATACTTGTAACAGTCATACATGTACATCATGTGCT
 405  A  N  Y  F  L  Y  M  G  G  C  Y  K  A  T  E  V  P  G  S  L  M  C  K
4279  GCTAATTACTTCTTATATATGGGTGGTTGTTATAAAGCAACAGAGGTGCCTGGTAGCCTCATGTGCAAG
 428  T  A  D  N  G  V  C  T  A  A  N  A  N  N  K  Y  F  V  V  P  G  A  T
4348  ACCGCTGACAACGGTGTCTGTACCGCTGCTAATGCCAACAATAAGTACTTCGTAGTACCAGGGGCAACC
 451  N  Q  N  Q  S  V  L  A  C  G  N  P  L  G  T  L  I  G  T  Q  G  T  A
4417  AATCAGAATCAGTCTGTGTTGGCCTGTGGTAACCCCCTAGGCACACTAATAGGCACTCAAGGTACTGCT
 474  K  H  T  L  G  M  D  G  C  S  Q  C  T  A  P  T  A  L  T  T  A  G  M
4486  AAGCATACGTTGGGAATGGATGGCTGCTCCCAGTGCACAGCCCCGACAGCTCTAACAACAGCTGGCATG
 497  T  A  A  I  C  T  A  C  D  S  G  K  K  P  N  R  D  G  S  G  C  V  L
4555  ACTGCTGCTATATGCACTGCCTGTGACAGTGGTAAGAAGCCCAACAGGGATGGCAGTGGGTGTGTGCTG
 520  C  S  V  G  D  C  K  S  C  V  M  D  N  I  C  G  E  C  N  S  G  F  S
4624  TGCTCTGTGGGTGATTGTAAGAGCTGTGTGATGGATAATATATGTGGGGAGTGCAACAGTGGCTTCAGT
 543  L  D  N  G  K  C  V  S  S  G  S  N  R  S  G  L  S  T  G  A  I  A  G
4693  CTAGACAATGGCAAGTGTGTGTCCTCTGGCAGTAACAGAAGCGGCCTCTCCACGGGGGCCATCGCGGGG
 566  I  S  V  A  V  I  A  V  V  G  G  L  I  G  F  L  C  W  W  F  V  C  R
4762  ATCTCCGTCGCTGTCATCGCTGTCGTGGGTGGTCTCATAGGGTTCCTCTGCTGGTGGTTCGTGTGTAGG
                                                                  *  *  W
AF4    G  K  A  *
4831  GGGAAGGCGTGACTTAGGTAGTGAATGCTCTATAGTGTATGATAATGGGTGATAGAGGGTGATAGTGG
   2  S  E  I  C  T  A  A  S  D  G  L  C  T  A  C  K  T  A  N  G  L  F  K
4900  AGTGAGATATGCACAGCAGCCAGTGATGGACTGTGCACTGCTTGTAAGACAGCTAACGGCCTGTTCAAG
  25  N  P  A  T  A  P  E  K  G  R  E  C  I  L  C  S  D  T  T  D  R  D  G
4969  AACCCTGCTACAGCACCAGAGAAGGGAAGGGAGTGCATCCTGTGCTCTGATACGACAGATAGAGACGGG
  48  V  T  G  A  A  G  C  S  E  C  S  H  T  G  T  S  G  P  A  T  C  T  V
5038  GTCACGGGAGCCGCTGGATGCTCAGAATGCTCCCACACGGGCACTAGCGGACCGGCCACATGCACTGTC
  71  C  Q  D  G  Y  I  K  K  G  D  A  C  E  K  C  D  Q  S  C  L  T  C  D
5107  TGCCAAGATGGATACATTAAGAAGGGTGATGCATGTGAGAAGTGTGATCAGTCTTGTCTGACCTGCGAT
  94  G  S  G  P  N  H  C  T  S  C  P  E  G  K  Y  L  K  T  D  K  S  C  V
5176  GGAAGTGGTCCGAATCATTGCACATCGTGTCCAGAGGGGAAGTACCTGAAGACTGACAAGTCATGCGTG
 117  N  N  N  G  C  T  G  N  T  Y  A  D  P  E  S  G  K  C  L  P  C  N  T
5245  AATAATAATGGATGCACTGGAAACACTTATGCTGACCCGGAGTCGGGGAAGTGCCTTCCTTGCAACACG
 140  I  D  Q  A  C  T  Q  C  E  V  D  S  T  T  K  K  P  K  C  T  A  C  D
5314  ATAGATCAAGCATGTACCCAGTGTGAGGTTGATTCGACTACCAAGAAGCCGAAGTGCACAGCTTGTGAT
 163  N  S  G  K  K  V  K  T  A  I  D  G  T  T  T  C  V  D  V  S
5383  AACAGTGGGAAAAAGGTAAAGACGGCCATCGATGGCACCACAACGTGCGTTGATGTGAGCTC
                                                                 Sac I
```

(**Fig. 4.15**) in order to construct a restriction map (**Fig. 4.16**). Hybridisation analyses of restriction digests indicated that no particular part of this insert hybridised better than the other segments with the M165-[5'] probe, i.e. fragments from the 3' and 5' ends, as well as those from the central portion, all hybridised strongly (data not presented). The insert was therefore sequenced from both ends as described earlier for pM7-1 and pM11-2 (sections 4.8.2a,b). Five sub-clones were constructed from restriction fragments obtained by cleaving pM13-3 DNA with single or paired endonucleases. These fragments were gel-purified and ligated with appropriately-cleaved pBluescript SK (+) (**Fig. 4.16**). The size of the subcloned fragments ranged from 0.8 to 3.5 kb. The flanks of each insert were sequenced using the T3 and T7 primers, which yielded overlapping sequence data, as outlined in **Fig. 4.16**. The complete 3,769-bp insert sequence of the pM13-3 clone is presented in **Fig. 4.17**. The fragment contained three ORF, designated *Sac*-D/1 ('DF1'), *Sac*-D/2 ('DF2') and *Sac*-D/3 ('DF3') (**Figs. 4.16, 4.17**). These are arranged in a head-to-tail tandem array. The DF1 reading frame was situated at the 5' end of the insert and because it overlapped the *Sac* I cleavage site, the 5' portion of the coding sequence was missing.

The DF1 reading frame was almost identical (99.3%) over its entire length (1,322-bp) to the corresponding 5' (*Sac* I) truncated BF1 (Sac-B/1) ORF identified in the pM11-3 construct, described in the next section. However, it is not known whether this similarity between the two ORF (DF1 of pM13-3, and BF1 of pM11-3) extends into their 5' segments because of the truncation within the respective inserts. Interestingly, the sequence similarity between the inserts of constructs B (pM11-3) and D (pM13-3) is limited to the BF1 and DF1 reading frames. Otherwise, no significant sequence homology was apparent between these two cloned genomic fragments (data not presented). The second (DF2) reading frame (1,728 bp) was found to have 62% nt sequence identity with the *crp*72 gene across the entire length of the ORF and 75% identity across an internal 1,000-bp segment (not shown). The third

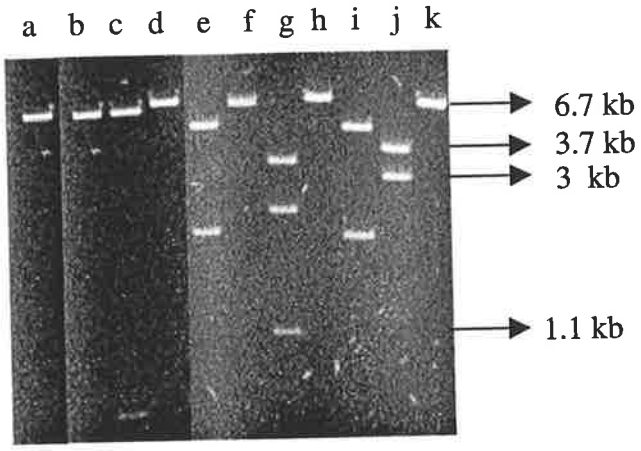


Figure 4.15. Restriction analysis of pM13-3 (construct D), containing a 3.8-kb genomic insert. The endonucleases used were:

a *Bam* H I
b *Cla* I
c *Dra* I
d *Eco* R I
e *Hind* III
f *Kpn* I
g *Pst* I
h *Sma* I
i *Sph* I
j *Xba* I
k *Xho* I

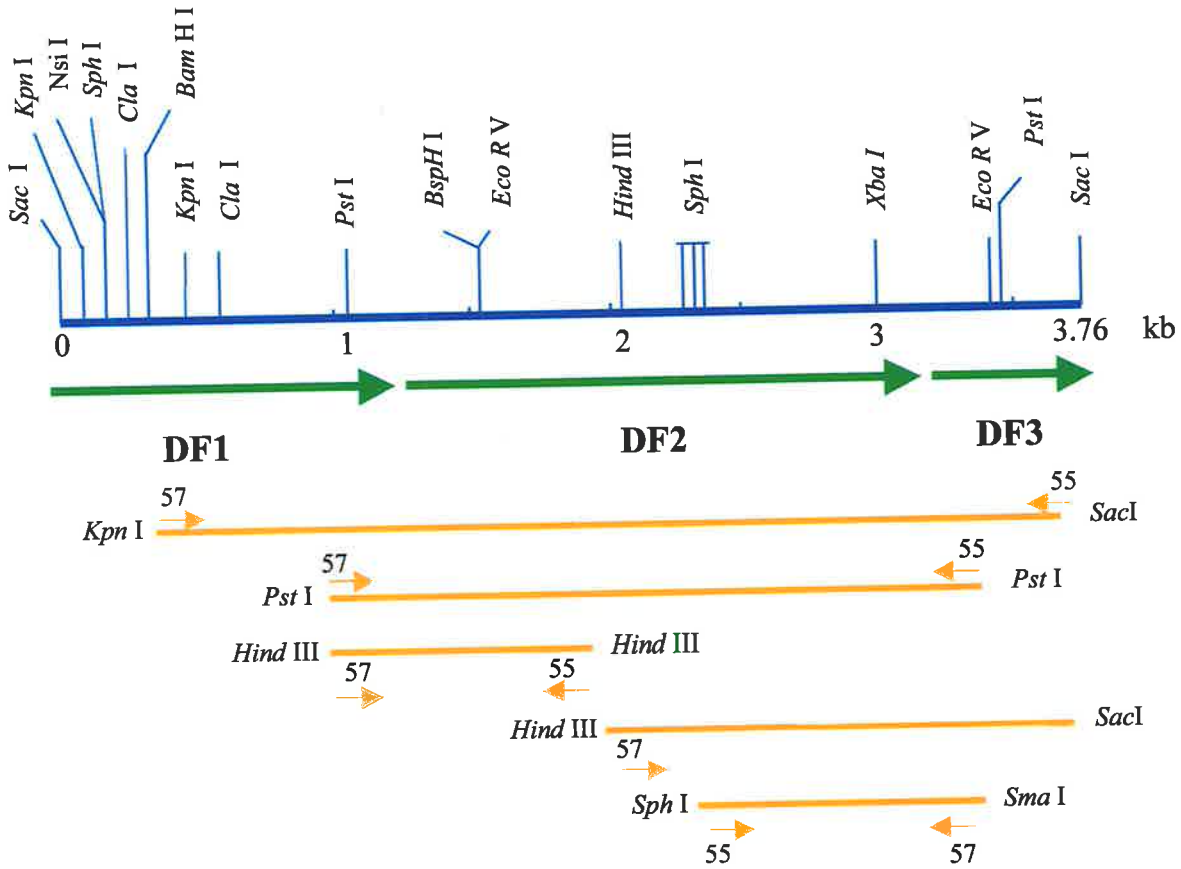


Figure 4.16. Schematic representation of the pM13-3 insert. Open reading frames DF1, DF2 and DF3 are denoted by thick arrows. Subcloned fragments (used for sequencing) are depicted by bold lines. Restriction sites that were identified experimentally by cleavage (and confirmed by sequence analysis) are shown. The identity and annealing position of primers (used for sequencing) are shown by short arrowheads.

Figure 4. 17. Nucleotide and deduced amino acid sequences specified by open reading frames within the 3,768-bp insert of pM13-3 (construct D). The three open reading frames (DF1, DF2, DF3) are indicated, together with encoded C-terminal invariant segment (CRGKA), and stop codons (both underlined), polyadenylation signal sequence (AGTRAA, double underlined) and VSP-specific pre polyA signal sequence (CTTAGRT, underlined). In frame stop codons that precede the open reading frames are highlighted. This sequence was submitted to the GeneBank™ database with accession number AF235029.

Sac I

DF1

```
   1  S  S  G  C  V  D  E  N  H  F  K  A  D  D  D  S  A  C  Y  L  C  G  D
   1 GAGCTCTGGATGTGTGGATGAGAATCACTTCAAGGCAGACGATGACTCAGCCTGTTATCTATGTGGTGAT
  23  I  S  G  N  N  E  N  T  N  K  G  V  A  D  C  N  K  C  T  K  E  A  G
  69 ATTAGTGGAAATAATGAGAATACAAACAAGGGTGTTGCCGATTGCAACAAATGCACCAAGGAGGCAGGT
  46  T  K  T  T  C  S  E  C  L  S  G  R  F  L  E  T  N  T  C  T  K  K  C
 138  CCAAAACAACGTGCTCTGAATGCCTATCGGGTCGTTTTCTTGAGACTAATACTTGTACGAAGAAATGC
  69  A  D  T  C  A  K  C  T  A  E  T  D  I  N  K  C  T  E  C  M  P  G  Y
 207 GCTGATACCTGTGCAAAGTGTACTGCTGAAACTGACATTAACAAATGTACTGAATGCATGCCGGGATAC
  92  F  L  K  T  E  G  V  T  K  E  C  V  P  C  S  D  T  A  K  G  G  I  D
 276 TTCCTAAAAACCGAGGGTGTAACAAAGGAGTGCGTCCCTTGTAGTGACACAGCCAAAGGTGGCATCGAT
 115  G  C  S  V  C  S  N  D  G  N  T  S  K  C  T  D  C  K  P  N  Y  R  K
 345 GGATGCAGTGTCTGCTCGAATGATGGCAATACTTCCAAGTGCACCGATTGTAAGCCCAACTATCGCAAG
 138  Q  S  D  S  S  P  G  S  V  T  C  T  K  T  C  E  D  P  T  A  C  G  G
 414 CAGAGCGATAGCAGTCCTGGCTCTGTCACATGTACCAAGACTTGTGAGGATCCTACAGCTTGTGGTGGT
 161  T  S  G  A  C  D  A  M  I  I  D  D  K  G  N  A  K  H  Y  C  S  Y  C
 483 ACCTCTGGAGCTTGCGATGCCATGATAATAGATGATAAGGGCAATGCAAAGCACTATTGTTCATACTGT
 184  G  E  T  N  K  F  P  I  D  G  I  C  N  S  D  A  Q  S  N  T  C  N  N
 552 GGAGAGACCAATAAGTTCCCTATCGATGGTATATGTAATAGTGATGCTCAAAGTAATACTTGTAACAAC
 207  H  V  C  T  S  C  T  T  G  Y  F  L  Y  M  G  G  C  Y  S  V  S  A  Q
 621 CATGTCTGTACATCATGTACTACGGGGTACTTCCTATACATGGGTGGCTGTTATAGTGTATCAGCTCAG
 230  P  G  N  Y  M  C  K  T  A  D  S  S  G  I  C  T  E  A  A  N  N  R  Y
 690 CCTGGCAACTATATGTGCAAGACAGCAGATAGTAGTGGGATATGCACAGAGGCGGCCAACAATAGGTAC
 253  F  L  V  P  G  A  S  N  T  D  Q  S  V  L  A  C  S  N  P  L  G  T  L
 759 TTCCTGGTCCCAGGGGCATCCAACACTGATCAGTCCGTGTTGGCTTGTAGCAACCCCCTAGGCACACTG
 276  T  G  T  G  D  T  A  K  A  Y  V  G  V  D  G  C  S  A  C  T  A  P  T
 828 ACAGGCACTGGTGATACTGCTAAGGCATATGTGGGAGTAGATGGCTGCTCAGCATGTACAGCCCCGACA
 299  Q  L  E  A  P  G  M  A  P  A  T  C  T  A  C  S  G  N  N  K  P  N  L
 897 CAGCTAGAGGCTCCTGGCATGGCCCCCGCCACGTGCACAGCGTGTAGCGGCAACAACAAGCCCAACCTG
 322  A  G  S  G  C  F  A  C  T  V  S  G  C  S  H  C  G  A  G  D  K  C  E
 966 GCTGGGAGCGGGTGCTTCGCATGTACCGTTAGTGGATGTTCGCACTGTGGAGCAGGCGATAAGTGCGAA
 345  G  C  T  S  K  D  Q  K  P  S  L  D  G  S  Q  C  I  A  C  S  I  D  G
1035 GGCTGTACCAGTAAAGACCAGAAGCCCAGCCTGGACGGCTCCCAGTGCATCGCCTGCAGTATCGACGGG
 368  C  V  R  C  S  E  E  N  K  C  G  Q  C  S  D  G  Y  R  L  E  G  G  R
1104 TGCGTCAGGTGCAGCGAGGAGAACAAGTGCGGCCAGTGCAGTGATGGGTACAGACTAGAGGGTGGCAGG
 391  C  V  S  S  G  A  N  R  S  G  L  S  A  G  A  I  A  G  I  S  V  A  V
1173 TGCGTGTCCTCTGGCGCCAACAGGAGCGGGCTCTCCGCGGGGGCCATCGCGGGGATCTCCGTGGCTGTG
 414  V  A  V  V  G  G  L  V  G  F  L  C  W  W  F  V  C  R  G  K  A  *  L
1242 GTCGCTGTCGTGGGGGGCCTCGTGGGCTTCCTCTGCTGGTGGTTCGTGTGTAGAGGGAAGGCGTGACTT
 437  *  *
1311 AGGTAGTGAATGCTATGAGTGATGATAATGGGTGATAATGTGTGATAGAGGGTGATAATGGGTGATAGC
```

DF2

```
                *  *  W  S  E  I  C  T  K  A  E  G  G  V  C  T  E  C  N
1380 CGTGTGATAGTATGATAGTGGAGTGAGATATGTACAAAGGCTGAGGGAGGAGTGTGCACTGAGTGTAAT
  18  T  A  N  G  L  F  K  N  P  A  A  T  P  E  K  G  S  E  C  I  L  C  H
1449 ACAGCTAACGGCCTGTTCAAGAACCCGGCTGCAACACCAGAGAAGGGTAGTGAGTGCATCCTCTGTCAT
  41  D  I  K  G  A  D  G  Y  T  G  V  A  N  C  A  Q  C  Q  A  P  Q  S  A
1518 GATATCAAGGGAGCAGATGGATATACGGGGGTGGCTAACTGCGCACAGTGTCAGGCACCACAGAGTGCA
  64  G  P  V  T  C  T  A  C  Q  D  G  F  V  K  K  D  S  A  C  V  K  C  G
1587 GGACCTGTCACATGTACTGCCTGTCAAGATGGGTTCGTTAAGAAGGACAGTGCATGTGTGAAGTGCGGA
  87  E  G  C  S  A  C  S  A  D  T  P  T  Q  C  T  A  C  V  E  G  K  F  L
1656 GAGGGCTGCTCTGCCTGCTCAGCTGATACTCCGACTCAGTGTACCGCCTGCGTTGAGGGGAAGTTCCTG
 110  K  A  D  K  C  V  D  A  N  Q  C  D  N  G  K  Y  A  D  P  K  T  G  Q
```

```
1725  AAGGCTGACAAGTGTGTGGATGCTAATCAATGTGACAATGGTAAGTATGCAGACCCAAAAACAGGCCAA
 133  C  K  A  C  T  D  T  S  V  N  E  C  A  T  C  A  Y  S  D  T  L  Q  K
1794  TGCAAGGCCTGCACGGACACAAGTGTCAATGAATGTGCTACGTGTGCATACAGCGACACTCTTCAGAAG
 156  P  V  C  T  G  C  N  S  G  G  N  L  L  L  K  V  N  P  D  G  S  A  T
1863  CCTGTGTGCACTGGGTGTAATAGTGGAGGAAACCTGCTTCTTAAGGTGAACCCCGATGGGTCAGCGACG
 179  C  V  A  E  A  E  C  T  S  G  N  T  H  F  L  E  Q  S  P  K  A  C  V
1932  TGTGTTGCGGAGGCAGAGTGCACAAGTGGCAATACGCACTTCCTTGAACAATCTCCAAAAGCTTGTGTT
 202  P  C  G  D  T  K  N  G  G  I  L  G  C  D  T  C  S  S  K  T  T  C  T
2001  CCATGTGGTGATACTAAAAATGGTGGGATTCTAGGTTGTGATACCTGCTCCTCTAAAACTACATGTACA
 225  K  C  L  D  G  Y  Y  N  S  A  S  G  N  T  A  T  C  T  A  C  G  E  N
2070  AAGTGCCTCGATGGATATTACAATAGTGCAAGTGGTAATACCGCTACATGCACAGCCTGTGGTGAGAAC
 248  C  A  T  C  A  K  D  T  K  D  Q  C  T  T  C  K  Q  G  Y  F  L  K  D
2139  TGTGCCACCTGTGCTAAGGATACAAAAGACCAATGCACGACCTGCAAGCAAGGGTACTTCTTGAAGGAT
 271  S  S  S  G  E  C  I  S  C  S  D  T  A  N  G  G  R  D  G  C  S  A  C
2208  AGCTCCTCTGGTGAGTGCATCTCATGTAGTGATACAGCAAATGGTGGCCGCGATGGGTGCAGTGCATGC
 294  S  N  S  G  G  F  K  C  T  D  C  K  P  N  Y  K  K  Q  L  N  G  D  A
2277  TCTAACAGTGGTGGATTCAAATGTACTGACTGTAAGCCTAACTATAAGAAACAGCTTAATGGAGATGCA
 317  G  D  D  Y  T  C  T  K  T  C  E  D  E  T  A  C  G  G  T  A  G  A  C
2346  GGGGATGACTATACGTGTACAAAGACCTGTGAAGATGAGACTGCATGCGGGGGGACAGCAGGGGCCTGC
 340  D  A  I  V  V  G  N  D  G  S  M  L  S  Y  C  S  K  C  V  G  A  G  Y
2415  GACGCCATCGTGGTCGGCAACGACGGCAGCATGCTCTCTTACTGTTCAAAGTGTGTTGGCGCCGGCTAT
 363  G  P  I  N  G  K  C  T  N  A  L  A  G  N  T  C  A  D  G  V  C  T  R
2484  GGCCCCATTAATGGGAAGTGCACCAATGCACTTGCAGGTAATACCTGCGCCGATGGTGTATGTACGCGG
 386  C  T  N  N  Y  F  L  Y  M  G  G  C  Y  K  A  T  E  V  P  G  S  L  M
2553  TGCACCAACAACTACTTCCTCTATATGGGTGGCTGTTATAAAGCAACAGAGGTGCCTGGTAGCCTCATG
 409  C  S  E  A  T  T  A  G  V  C  T  T  P  N  A  N  D  R  Y  F  V  V  P
2622  TGCTCCGAAGCAACAACAGCTGGTGTCTGTACTACTCCCAATGCCAATGATAGGTACTTCGTAGTCCCA
 432  E  A  K  A  T  D  Q  S  V  L  A  C  G  N  P  L  G  T  L  V  D  P  Q
2691  GAAGCAAAGGCTACTGATCAGTCTGTACTAGCCTGTGGCAACCCCCTAGGCACACTAGTAGACCCTCAA
 455  G  T  A  K  A  Y  V  G  V  D  G  C  S  Q  C  T  A  P  T  A  L  T  E
2760  GGTACTGCTAAGGCATACGTGGGGGTGGATGGCTGCTCACAGTGCACAGCCCCAACAGCTCTGACTGAG
 478  G  G  M  A  A  A  V  C  T  S  C  D  S  D  R  K  P  N  R  D  G  S  G
2829  GGTGGTATGGCTGCTGCTGTGTGTACCTCCTGTGACAGTGATAGGAAGCCCAACAGGGACGGCAGTGGG
 501  C  V  L  C  S  V  G  G  C  K  S  C  V  V  D  G  V  C  G  E  C  N  S
2898  TGTGTCCTGTGCTCTGTGGGTGGTTGTAAGAGCTGTGTGGTGGACGGTGTGTGTGGGGAGTGCAACAGT
 524  G  F  S  L  D  N  G  K  C  V  S  S  G  A  N  R  S  G  L  S  T  G  A
2967  GGCTTCAGTCTAGACAATGGTAAGTGTGTGTCCTCTGGCGCCAACAGGAGCGGGCTCTCCACGGGGGCC
 547  I  A  G  I  S  V  A  V  I  A  V  V  G  G  L  V  G  F  L  C  W  W  F
3036  ATCGCGGGGATCTCCGTCGCTGTCATCGCTGTCGTGGGGGGCCTCGTGGGCTTCCTCTGCTGGTGGTTC
 570  V  C  R  G  K  A  *
3105  GTGTGTAGAGGGAAGGCGTGACTTAGGTAGTGAATGCTATGAGTGATGATAATGGGTGATAATGTGTG-
DF3                                                  *  *  W  S  E  I  C  T  K  A  E
3174  ATAGAGGGTGATAATGGGTGATAGTGTGTGATAGTATGATAGTGGAGTGAGATATGTACAAAGGCTGAG
  10  G  G  V  C  T  E  C  N  T  A  N  G  L  F  K  N  P  A  A  T  P  E  K
3243  GGAGGAGTGTGCACTGAGTGTAATACAGCTAACGGCCTGTTCAAGAACCCGGCTGCAACACCAGAGAAG
  33  G  R  E  C  I  L  C  S  D  I  N  G  A  D  G  Y  T  G  V  A  N  C  A
3312  GGAAGGGAGTGCATACTGTGCTCCGATATCAATGGAGCAGATGGATATACGGGGGTGGCTAACTGCGCA
  56  Q  C  T  K  S  D  S  S  T  G  P  A  T  C  S  T  C  R  D  G  Y  Y  M
3381  CAGTGTACCAAGTCAGACAGTAGTACAGGACCTGCCACGTGTAGTACCTGCAGAGATGGATACTACATG
  79  D  S  Q  A  C  T  K  C  N  D  N  C  A  T  C  T  G  A  G  Q  N  Q  C
3450  GACTCTCAGGCATGTACCAAATGTAATGACAACTGTGCCACATGTACCGGGGCCGGTCAGAATCAGTGT
 102  S  S  C  K  A  G  F  Y  L  K  S  D  G  S  C  S  K  T  C  D  N  N  Q
3519  TCATCTTGTAAAGCGGGCTTCTATCTGAAGTCAGACGGGTCGTGCTCAAAAACCTGCGATAATAACCAG
 125  Y  P  D  P  S  T  G  K  C  T  A  C  N  G  A  A  E  Q  G  G  I  P  E
3588  TACCCCGACCCAAGCACCGGCAAGTGCACAGCCTGTAACGGTGCAGCAGAGCAGGGGGGCATACCAGAG
 148  C  T  A  C  T  Y  D  A  K  L  Q  K  P  V  C  S  D  C  G  N  K  K  V
3657  TGCACTGCTTGCACGTACGATGCAAAGCTACAGAAGCCGGTATGCAGTGACTGCGGTAATAAGAAGGTA
 171  K  T  E  L  D  G  T  T  T  S  V  H  M  S
3726  AAGACAGAGCTGGATGGAACGACAACCTCCGTCCATATGAGCTC
                                                  Sac I
```

(DF3) reading frame comprised the 553 bp at the 3' end of the insert. It was truncated by the *Sac* I cloning site (**Fig. 4. 17**).

4.8.2d Construct B (pM11-3)

The pM11-3 construct (construct B) contained a 5.5-kb *Sac* I insert. This was subjected to a detailed restriction mapping analysis, supplemented by Southern hybridisations using the M165-[5'] probe as exemplified by the data presented in **Fig. 4.18** and **Table 4.6**. A 3.5 kb *Hin*d III fragment, derived from the 5' portion of the insert, hybridised strongly with the probe (**Fig. 4.18B**). This fragment was subcloned into pBluescript SK(+) and preliminary sequence data were obtained from both ends. Two additional subclones containing the 1,812-bp *Kpn* I fragment and the 1,172-bp *Kpn* I / *Hind* III fragment (**Figs. 4.18, 4.19**) were also constructed and used for flanking sequence determinations as indicated in **Fig. 4.19**. Additional sequence data were obtained using primers 180, 181 and 182, designed on the basis of the previously acquired partial sequences. The complete nucleotide sequence of the subcloned 3,475-bp *Hin*d III fragment is presented in **Fig. 4.20**. Analysis of this sequence for ORF identified two vsp gene sequences, designated Sac-B/1 ('BF1') and Sac-B/2 ('BF2') (**Figs. 4.19, 4.20**). The BF1 reading frame (1,309 bp) was incomplete, as it was truncated at the 5' end by the cleavage at the *Sac* I cloning site.

The BF2 reading frame commenced in a head-to-tail arrangement, 21 bp beyond the Sac-B/1 polyadenylation signal sequence (**Fig. 4.20**). These two reading frames exhibited nt identity at 76.3% of sites over a 539-bp overlap within their 3'regions, but a much lower level of identity (58.7%) over a 1,231-bp overlap in their upstream regions. Comparisons of the BF1 and BF2 ORFs with the *crp*72 gene by FASTA analysis revealed only 63% nt identity over 1,196-bp and 1,386-bp overlaps respectively, although a much higher level of nt identity with *crp*72 was observed if only the 500-bp at the 3' end were compared (80% for BF1, 90% for BF2). Elucidation of the nucleotide sequence of an additional 447-bp segment

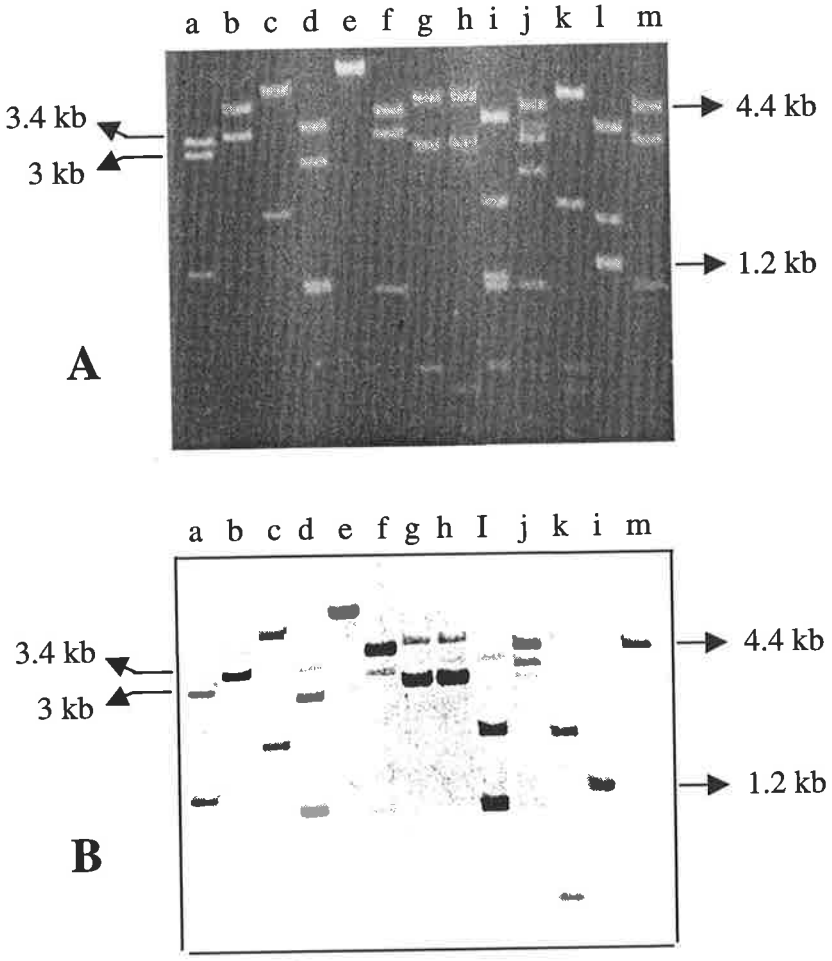


Figure 4.18. Restriction and Southern hybridisation analysis of pM11-3 (constructB). Replicate aliquots of plasmid DNA were incubated with different restriction endonucleases. The reaction mixtures were subjected to electrophoresis on 1% agarose and stained with ethidium bromide (A). The gel was subsequently analysed by Southern hybridisation (B) using the DIG-labelled M165-[5'] probe. Size markers are indicated. The enzymes used were :

a	*Cla* I	h	*Hind* III + *Xba* I
b	*Hind* III	i	*Pst* I + *Xba* I
c	*Kpn* I	j	*Sph* I + *EcoR* I
d	*Pst* I	k	*Kpn* I + *Xba* I
e	*Sma* I	l	*Hind* III + *Cla* I
f	*Sph* I	m	*Sma* I + *Sph* I
g	*Xba* I		

Table 4.6. Restriction endonuclease cleavage data and associated Southern hybridisation data for plasmid construct pM11-3 (construct 'B')

A. *Single-enzyme cleavage tests*

Enzyme tested	No. of fragments	No. of sites in vector	No. of sites in insert	Fragment sizes (kb)	Sum of fragments (kb)	Identity* of fragment (site1-site2)	Hybridisation with probe 154-155
Bam H1	1	1	0				
Dra I	2	2	0				
*Eco*R1	1	1	0				
Sma I	1	1	0				
Xho I	1	1	0				
Cla I	(6) Observed 4	1	5	3.40		C5 - Co	-
				3.00		C4 - C5	+++
				1.20		C2 - C3	+++
				0.37		C3 - C4	-
				0.34		Co - C1	-
				0.23	8.5	C1 - C2	-
*Hin*d III	2	1	1	5.00		H1 - Ho	-
				3.40	8.4	Ho - H1	+++
Kpn I	(4) Observed 3	1	3	6.20		K3 - Ko	+++
				1.80		K2 - K3	+++
				0.35		K1 - K2	-
				0.14	8.5	Ko - K1	-
Pst I	4	1	3	3.70		P3 - Po	-
				2.70		P1 - P2	+++
				1.16		P2 - P3	-
				1.14	8.7	Po - P1	+++
Sph I	3	0	3	4.20		S1 - S2	+++
				3.40		S3 - S1	-
				1.00	8.6	S2 - S3	-
Xba I	3	1	2	4.90		X2 - Xo	-
				2.90		Xo - X1	++++
				0.72	8.5	X1 - X2	-

Average sum of fragments: 8.54 kb
***less* Size of vector:** 2.95 kb
Calculated size of cloned insert: 5.59 **kb**

B. *Double-enzyme cleavage tests*

							Hybridisation with probe 154-155
*Hin*d III + *Xba* I	(4) Observed 3	1 + 1	1 + 2	4.90		X2 - Xo	-
				2.90		Xo - X1	+++
				0.65		X1 - H1	-
				0.10	8.6	H1 - X2	-
Kpn I + *Xba* I	(7) Observed 4	1 + 1	3 + 2	4.90		X2 - Xo	-
				1.80		K2 - K3	+++
				0.70		X2 - X3	-
				0.60	8.0	K3 - X2	+++
Pst I + *Xba* I	(6) Observed 5	1 + 1	3 + 2	3.60		P3 - Po	-
				1.80		P1 - X1	+++
				1.16		P2 - P3	-
				1.10		Po - P1	+++
				0.70		X1 - X2	-
				0.70	9.1	X2 - P2	-
*Hin*d III + *Cla* I	(7) Observed 6	1 + 1	1 + 5	3.50		C5 - Co	-
				1.70		H1 - C5	-
				1.25		C4 - H1	-
				1.25		C2 - C3	+++
				0.34		Co - C1	-
				0.30	8.3	C1 - C2	-
Sma I + *Sph* I	(4) Observed 3	1 + 0	0 + 4	4.20		S1 - S2	+++
				3.10		S3 - Sma	-
				1.00		S2 - S3	-
				0.27	8.6	Sma - S1	-

Average sum of fragments: 8.50
***less* Size of vector:** 2.95
Calculated size of cloned insert: 5.55

*Cleavage sites are designated by the first letter of the relevant enzyme, followed by the order of the site(s) in the cloned insert, e.g. C_5 - C_0 represents the fragment spanning *Cla* I site (5) to the 3' end of the insert (cleaved by the *Cla* I site within the multiple cloning site of the vector. Similarly, P2 - P3 represents the fragment spanning Pst I sites (1) and (2).

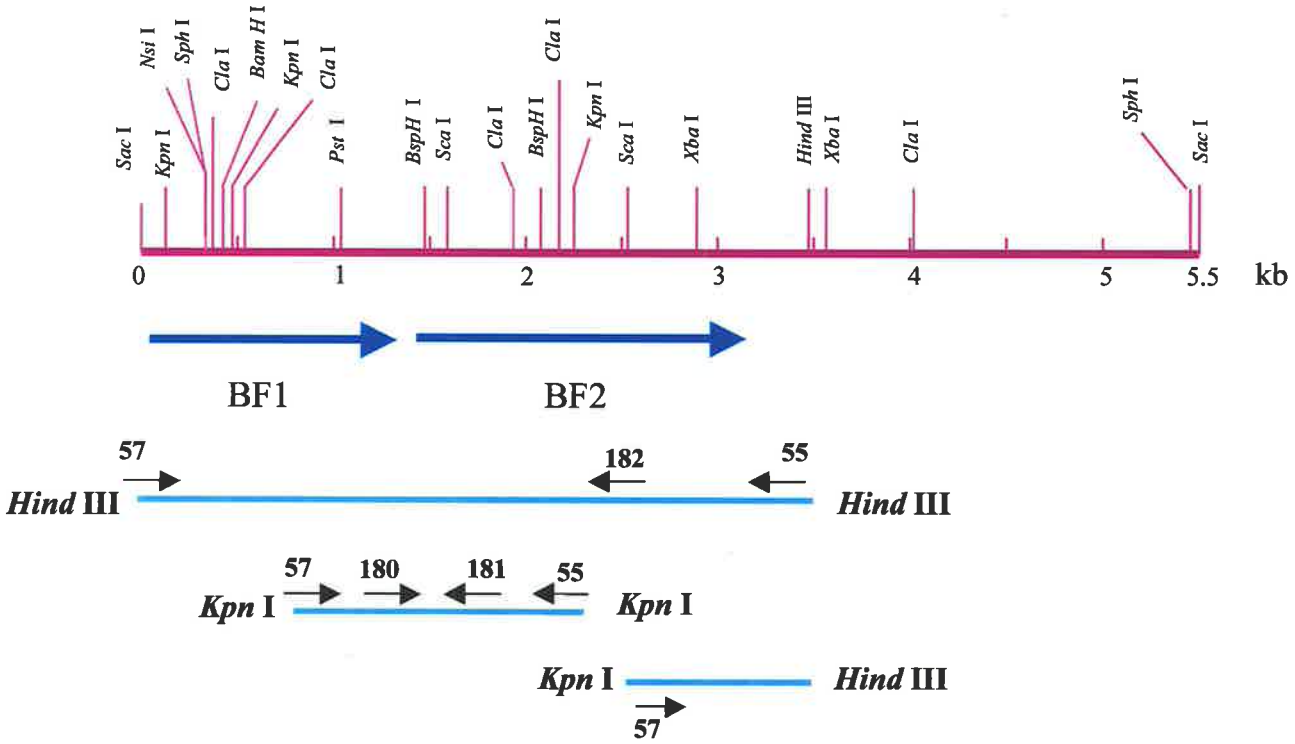


Figure 4.19. Schematic representation of the pM11-3 insert. Open reading frames BF1 and BF2 are denoted by thick arrows. Subcloned fragments used for sequencing are depicted by bold lines. Restriction sites that were identified experimentally by cleavage are shown. The identity and annealing position of primers (used for sequencing) are shown by short arrowheads.

Figure 4.20. Nucleotide and deduced amino acid sequences specified by open reading frames within the 3,475-bp insert of pM11-3 (construct B). The two open reading frames BF1 and BF2 are indicated, together with encoded C-terminal invariant segment 'CRGKA' and stop codons (both underlined), polyadenylation signal sequence (AGTRAA, double underlined) and the VSP-specific pre-polyA signal sequence (CTTAGRT, underlined). In-frame stop codons that precede the BF2 open reading frame are highlighted. This sequence was submitted to the Genebank™ database with accession number AF235028.

Sac I

BF1

```
      S  S  G  C  V  D  E  N  H  F  K  A  D  D  D  S  A  C  Y  L  C  G
   1  GAGCTCTGGATGTGTGGATGAGAATCACTTCAAGGCAGACGATGACTCAGCCTGTTATCTATGTGGT
  23  D  I  S  G  N  N  E  N  T  N  K  G  V  A  D  C  N  K  C  T  K  E  A
  68  GATATTAGTGGAAATAATGAGAATACAAACAAGGGTGTTGCCGATTGCAACAAATGCACCAAGGAGGCA
  46  G  T  K  T  T  C  S  E  C  L  S  G  R  F  L  E  T  N  T  C  T  K  K
 137  GGTACCAAAACAACGTGCTCTGAATGCCTATCGGGTCGTTTTCTTGAGACTAATACTTGTACGAAGAAA
  69  C  A  D  T  C  A  K  C  T  A  E  T  D  I  N  K  C  T  E  C  M  P  G
 206  TGCGCTGATACCTGTGCAAAGTGTACTGCTGAAACTGACATTAACAAATGTACTGAATGCATGCCGGGA
  92  Y  F  L  K  T  E  G  V  T  K  E  C  V  P  C  S  D  T  A  K  G  G  I
 275  TACTTCCTAAAAACCGAGGGTGTAACAAAGGAGTGCGTCCCTTGTAGTGACACAGCCAAAGGTGGCATC
 115  D  G  C  S  V  C  S  N  D  G  N  T  S  K  C  T  D  C  K  P  N  Y  R
 344  GATGGATGCAGTGTCTGCTCGAATGATGGCAATACTTCCAAGTGCACCGATTGTAAGCCCAACTATCGC
 138  K  Q  S  D  S  S  P  G  S  V  T  C  T  K  T  C  E  D  P  T  A  C  G
 413  AAGCAGAGCGATAGCAGTCCTGGCTCTGTCACATGTACCAAGACTTGTGAGGATCCTACAGCTTGTGGT
 161  G  T  S  G  A  C  D  A  M  I  I  D  D  K  G  N  A  K  H  Y  C  S  Y
 482  GGTACCTCTGGAGCTTGCGATGCCATGATAATAGATGATAAGGGCAATGCAAAGCACTATTGTTCATAC
 184  C  G  E  T  N  K  F  P  I  D  G  I  C  N  S  D  A  Q  S  N  T  C  N
 551  TGTGGAGAGACCAATAAGTTCCCTATCGATGGTATATGTAATAGTGATGCTCAAAGTAATACTTGTAAC
 207  N  H  V  C  T  S  C  T  T  G  Y  F  L  Y  M  G  G  C  Y  S  V  S  A
 620  AACCATGTCTGTACATCATGTACTACGGGGTACTTCCTATACATGGGTGGCTGTTATAGTGTATCAGCT
 230  Q  P  G  N  Y  M  C  K  T  A  D  S  S  G  I  C  T  E  A  A  N  N  R
 689  CAGCCTGGCAACTATATGTGCAAGACAGCAGATAGTAGTGGGATATGCACAGAGGCGGCCAACAATAGG
 253  Y  F  L  V  P  G  A  S  N  T  D  Q  S  V  L  A  C  S  N  P  L  G  T
 758  TACTTCCTGGTCCCAGGGGCATCCAACACTGATCAGTCCGTGTTGGCTTGTAGCAACCCCCTAGGCACA
 276  L  T  G  T  G  D  T  A  K  A  Y  V  G  V  Y  G  C  S  A  C  T  A  P
 827  CTGACAGGCACTGGTGATACTGCTAAGGCATATGTGGGAGTATATGGCTGCTCAGCATGTACAGCCCCG
 299  T  Q  L  E  A  P  G  M  A  P  A  T  C  T  A  C  I  G  N  N  K  P  N
 896  ACACAGCTAGAAGCTCCTGGCATGGCCCCCGCCACGTGCACAGCGTGTATCGGCAACAACAAGCCCAAC
 322  L  A  G  S  G  C  F  A  C  T  V  S  G  C  S  H  C  G  A  G  D  K  C
 965  CTGGCTGGGAGCGGGTGCTTCGCATGTACCGTTAGTGGATGTTCGCACTGTGGAGCAGGCGATAAGTGC
 345  E  G  C  T  S  K  D  Q  R  P  S  L  D  G  S  Q  C  I  A  C  S  I  D
1034  GAGGGCTGTACCAGTAAAGACCAGAGGCCCAGCCTGGACGGCTCCCAGTGCATCGCCTGCAGTATCGAC
 368  G  C  V  R  C  S  E  E  N  K  C  G  Q  C  S  D  G  Y  R  L  E  G  G
1103  GGGTGCGTCAGGTGCAGCGAGGAGAACAAGTGCGGCCAGTGCAGTGATGGGTACAGACTAGAGGGTGGC
 391  R  C  V  S  S  G  A  N  R  S  G  L  S  A  G  A  I  A  G  I  S  V  A
1172  AGGTGCGTGTCCTCTGGCGCCAACAGGAGCGGGCTCTCCGCGGGGGCCATCGCGGGGATCTCCGTGGCT
 414  V  V  A  V  V  G  G  L  V  G  F  L  C  W  W  F  V  C  R  G  K  A  *
1241  GTGGTCGCCGTCGTGGGGGGCCTCGTCGGCTTCCTCTGCTGGTGGTTCGTGTGTAGAGGGAAGGCGTGA
```

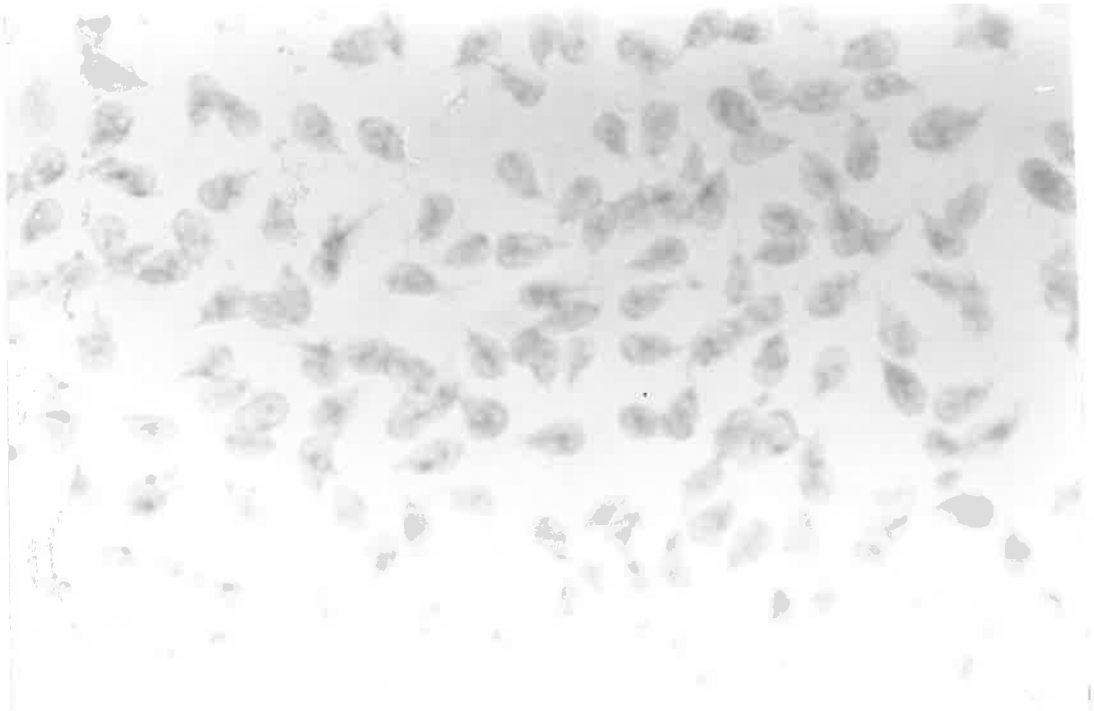
BF2

```
                                      *  *  W  V  I  C  T  A  A  S  N  G  L  C
1310  CTTAAGTAGTGAATGCTGTGCAGTGTGTGATAGTGGGTGATATGCACAGCAGCCAGTAATGGACTGTGC
  13  C  A  C  K  T  D  N  K  Y  V  F  Q  N  P  A  T  A  P  E  K  G  R  E
1379  ACTGCTTGTAAGACAGATAACAAGTACGTATTCCAGAACCCTGCTACAGCACCAGAGAAGGGAAGGGAG
  36  C  I  L  C  H  D  A  T  G  A  D  G  Y  T  G  V  A  N  C  A  T  C  T
1448  TGCATACTCTGTCATGATGCCACGGGAGCAGATGGGTATACGGGGGTGGCTAACTGCGCCACATGCACC
  59  K  S  D  S  N  K  G  A  A  T  C  T  A  C  Q  A  G  Y  Y  K  D  F  Q
1517  AAGTCAGATAGTAATAAGGGGGCAGCCACGTGTACTGCCTGTCAAGCAGGGTACTACAAGGACTTTCAG
  82  A  C  S  K  C  D  G  T  C  L  T  C  E  T  S  A  A  Q  C  T  S  C  P
1586  GCATGTAGCAAGTGTGATGGTACTTGTCTGACCTGTGAGACATCGGCGGCCCAGTGCACATCGTGTCCA
 105  E  G  K  Y  L  K  G  D  K  T  C  V  T  S  D  Q  C  T  S  T  V  Y  P
1655  GAGGGGAAGTACTTGAAGGGTGATAAGACTTGCGTAACTAGTGATCAATGCACGAGCACCGTTTATCCA
 128  D  P  E  S  G  T  C  K  E  C  S  T  I  D  Q  A  C  T  T  C  K  Y  N
1724  GACCCAGAGTCTGGAACGTGCAAGGAATGCAGTACGATAGATCAAGCATGTACTACCTGTAAGTACAAT
```

```
 151  A  T  V  S  K  P  Q  C  T  A  C  N  N  N  K  K  V  K  T  A  I  D  G
1793  GCGACTGTCAGTAAGCCCCAGTGCACCGCCTGTAATAACAATAAGAAGGTGAAGACGGCCATCGATGGC
 174  T  T  T  C  V  D  I  S  S  G  C  E  D  A  D  H  F  K  A  D  G  D  A
1862  ACCACGACATGCGTCGATATAAGCTCTGGCTGCGAGGATGCTGATCACTTCAAGGCAGATGGTGATGCT
 191  A  C  V  L  C  S  D  T  S  G  S  D  P  K  N  K  G  I  A  G  C  K  A
1931  GCGTGTGTCTTGTGCAGCGATACCAGCGGATCAGACCCAAAGAACAAAGGCATTGCTGGCTGTAAGGCT
 220  C  T  K  T  A  S  A  N  P  T  C  S  E  C  L  A  G  Y  Y  S  S  G  S
2000  TGCACGAAGACAGCAAGTGCCAATCCAACCTGCTCAGAGTGCTTAGCTGGATACTACAGTAGTGGTAGT
 243  G  T  V  T  C  A  A  C  G  D  N  C  A  T  C  T  Q  T  G  N  N  Q  C
2069  GGTACGGTTACATGCGCAGCCTGTGGTGATAATTGTGCCACCTGCACTCAAACTGGCAATAATCAGTGT
 266  T  K  C  K  P  G  F  F  M  K  G  S  G  P  T  G  E  C  V  A  C  D  N
2138  ACTAAATGCAAGCCTGGATTCTTCATGAAAGGCAGTGGTCCTACTGGTGAGTGCGTTGCTTGTGATAAT
 289  A  Q  G  G  I  D  G  C  A  T  C  T  F  S  G  S  L  T  C  N  S  C  K
2207  GCACAGGGGGGCATCGATGGATGTGCAACGTGCACCTTCTCTGGCTCGTTGACCTGCAACTCCTGTAAG
 312  P  N  Y  K  Q  N  G  T  S  G  A  C  D  A  M  I  I  D  D  Q  G  T  T
2276  CCCAACTACAAGCAAAACGGTACCTCTGGCGCCTGTGATGCCATGATAATAGATGATCAGGGCACCACA
 335  K  H  Y  C  S  Y  C  G  E  A  N  K  F  P  I  D  G  I  C  T  S  D  N
2345  AAGCACTATTGTTCTTACTGTGGAGAGGCTAATAAGTTCCCTATTGATGGTATATGTACTAGTGATAAT
 358  N  K  K  G  T  N  T  C  D  S  H  T  C  T  Q  C  A  Q  G  Y  F  L  Y
2414  AATAAGAAAGGAACTAATACCTGTGACAGTCATACATGTACACAGTGTGCTCAAGGGTACTTCTTATAT
 381  M  G  G  C  Y  K  V  G  Q  E  P  G  S  F  M  C  K  A  S  S  S  N  G
2483  ATGGGAGGCTGTTATAAGGTTGGTCAAGAACCAGGTAGCTTCATGTGCAAGGCATCAAGTAGTAATGGG
 404  I  C  T  E  A  A  S  N  K  Y  F  L  V  P  G  A  S  N  T  D  Q  S  V
2552  ATATGCACAGAGGCAGCAAGTAATAAGTACTTCCTGGTCCCAGGGGCATCCAACACTGATCAGTCCGTG
 427  L  A  C  S  N  P  L  G  T  L  T  G  T  G  D  T  A  K  A  Y  V  G  V
2621  TTGGCTTGTAGCAACCCCCTAGGCACACTGACAGGCACTGGTGATACTGCTAAGGCATACGTGGGGGTG
 450  D  G  C  S  Q  C  T  A  P  T  A  L  T  E  G  G  M  A  A  A  V  C  T
2690  GATGGCTGCTCACAGTGCACAGCCCCAACAGCTCTGACTGAGGGTGGTATGGCTGCTGCTGTGTGTACC
 473  S  C  D  S  D  R  K  P  N  R  D  G  S  G  C  V  L  C  S  V  G  G  C
2759  TCCTGTGACAGTGATAGGAAGCCCAACAGGGACGGCAGTGGGTGTGTCCTGTGCTCTGTGGGTGGTTGT
 496  K  N  C  V  M  D  N  I  C  G  E  C  N  S  G  F  S  L  D  N  G  K  C
2828  AAGAACTGTGTGATGGATAATATATGTGGGGAGTGCAACAGTGGCTTCAGTCTAGACAATGGTAAGTGT
 519  V  S  S  G  T  N  R  S  G  L  S  T  G  A  I  A  G  I  S  V  A  V  V
2897  GTGTCCTCTGGCACCAACAGGAGCGGCCTCAGCACAGGCGCCATCGCGGGGATCTCCGTCGCCGTGGTC
 542  V  V  V  G  G  L  V  G  F  F  C  W  W  F  I  C  R  G  K  A  *  L  R
2966  GTCGTCGTGGGGGGCCTCGTCGGCTTCTTCTGCTGGTGGTTCATATGTAGGGGGAAGGCGTGACTCAGG
1012  *  *
3035  TAGTAAACGCGTCCACCGTTATGTAACCTCACTATGCTCTCATACAATCTGCCTGGATAAGAGGAAAAG
3104  TCGCCGACAAGCCTCAGCTGTGCTCTGTTCAGGACCACCACACAGAGGCGAGAAGTCGGCTGGCCGTCA
3173  TTTTTTAGCAGTGACATCGACTGCTGCCTCTACTGGGTCCTCCTGGAATGCNTAGAGCGCCCACACATG
3242  GATTNCGTATGTACCCTCCACGGCTATCTGGGCATGATCGCCACAGCTTCTCGTAACCCTGTTTCCCCT
3311  GCCACAGGTTCCATGCGCATCTGCTCATAAGAGGATAACACTCACTGTGACCAGAATCAGAAGTGTCGA
3380  TGGCCCACGTTCCGGGGCTCAACCGTGCAGATTGTGTGCCCAGTCCGTTCAGCACTTACGGACAGTGCC
3449  TCTGATAGTGGACTTATAACTAAGCTT
```

***Hin*d III**

a



downstream from the BF2 reading frame revealed no additional ORF. This result was consistent with data from the Southern hybridisations, which had shown that only the 3.47-kb *Hind* III fragment (located at the 5' portion of the parent insert) hybridised with the M165-[5]' probe (**Fig. 4.18, Table 4.6**).

4.8.2e Construct E (pM3-3E)

A colony containing this construct was identified by screening the first *Giardia* (Ad-1/c3) *Sac* I genomic library with the M165-[3'] probe. Southern hybridisation analysis revealed a *Hind* III fragment, derived from the central portion of the 8.4-kb insert, that hybridised very strongly with this probe (data summarised in **Tables 4.7** and **4.8**). Cleavage with other restriction endonucleases produced multiple fragments that hybridised with the probe. Subclones containing the ≈ 1.95 kb *Hind* III and ≈ 3 kb *Pst* I fragments (**Fig. 4.21**) were produced by ligation with pBluescript SK(+). These constructs were used to obtain preliminary (flanking) sequence information, using T3 and T7 primers. Additional primers were designed on an ongoing basis from these initial and additional, accumulated sequence data obtained during the analysis of the two subclones (**Fig. 4.21**). The compiled data, representing a 2,792-bp segment spanning the two subclones (**Fig. 4.21**), as presented in **Fig. 4.22**.

Two ORF were identified within this sequence. The first (EF1) was located at the 5' end of the sequence. It was truncated by the *Pst* I cleavage site which was used to subclone the *Pst* I fragment (**Fig. 4.21**). The second ORF (EF2; 1,775-bp) commenced 83 bp beyond the EF1 stop codon, in a head-to-tail orientation and a –1 frameshift relative to EF1 (**Figs. 4.21, 4.22**). Both ORF encoded typical VSP amino acid sequences with conserved C-terminal segments. Each also possessed an intact stop codon and 3' extended polyadenylation signal sequence. The EF1 coding sequence ended with a TGA codon, 6 bp beyond which (i.e. in the same frame) were two additional adjacent stop codons (TAGTGA, part of the extended

Table 4.8. Restriction endonuclease cleavage data and associated Southern hybridisation data for plasmid construct pM3-3e (construct E)

Double - enzyme cleavage

Enzymes tested	No. of fragments	Sites in vector	Sites in insert	Fragment sizes (kb)	Sum of fragments (kb)	Identity * of fragment (site1-site2)	Hybridisation with probe A 148-149
*Bam*H1+ *Xba* I	3	1 + 1		7.50		X - B	
				3.70		B - V	
				0.72	11.9	X - X	
Cla I + *Dra* I	4	1 + (3)	0 + 1?	8.50		D - VD1	
				1.80		Co - D	
				1.00		VD3 - Co	
				0.72	12.0	VD1/2-VD3	
Cla I + *Sca* I	1	1 + 0	Linear.				
Pst I + *Xba* I	4	1 + 1		6.50		P2 - Vo	-
				3.03		P1 - P2	++
				1.61		Xb - P1	-
				0.69	11.8	Vo - Xb	-
Pst I + *Eco* R1	5	1 + 1		3.50		E2 - Vo	+
				3.00		P1 - P2	+++
				2.80		P2 - E2	+
				1.61		E1 - P1	-
				0.66	11.6	Eo - E1	-
Pst I + *Sph* I	6	1 + 0	2 + 4	3.60		S4 - Vo	+/-
				2.75		P2 - S4	-
				2.00		S1 - P1	-
				1.35		S2 - S3	-
				1.25		P1 - S2	++++
				0.56	11.5	Po - S1	-
Bam H1 + *Eco* R1	3	1 + 1	1 + 2	7.40		E1 - B/E2	++
				3.50		B/E2 - Vo	+/-
				0.66	11.6	Vo - E1	-
Xba I + *Eco* R1	3	1 + 1		7.40		E1/Xb - E2	+++
				3.50		E2 - Vo	+/-
				0.62	11.5	XE - EX	-
Pst I + *Hin*d III	7	1 + 1		6.10		P3 - Vo	++
				1.30		P1 - H4	+++++
				1.10			-
				0.72			-
				0.70			-
				0.55			-
				0.40	10.9		-
Sph I + *Hin*d III	7	0 + 1		3.50		S4 - S5	++++
				3.00		S5 - Vo	-
				1.95		H2 - S2	++++
				0.75			-
				0.70			-
				0.40			-
				0.33	10.6		-
Sph I + *Eco*R1	5	0 + 1		3.50		S4 - S5	+
				3.20		S5/E2 - Vo	-
				2.90		E1 - S2	+++++
				1.30		S2 - S3	-
				0.40	11.3	Eo - E1	-
Dra I + *Hin*d III	7	(3) + 1	(1) + 4	5.40		H4 - VD1	-
				1.90		D1 - H3	+++++
				1.10		VD3 - Ho	-
				1.00		Ho - H1/2	-
				0.76			-
				0.70		VD1/2-VD3	+
				0.48	11.3		-
Dra I + *Pst* I	6	(3) + 1	(1)+2	4.60		D4 - Vo	-
				3.00		P1 - P2	+++++
				1.70		Po - D1	-
				1.10		VD3 - Po	-
				0.72		VD1/2-VD3	+
				0.48	11.6	D1 - P1	-

Average sum of fragments:	11.47	kb
***less* Size of vector:**	2.95	kb
Calculated size of cloned insert:	**8.52**	**kb**

* Cleavage sites are designated by the first letter of the relevant enzyme, followed by the order of the site(s) in the cloned insert, e.g. P_2 - V_0 represents the fragment spanning *Pst* I site (2) to the 3' end of the insert, which was cleaved at *Pst* I site (2) and the *Pst* I site (0, V_0) within the multiple cloning site of the vector. Similarly, P_2 - E_2 represents the fragment spanning *Pst* I site (2) and *Eco*R1 site (2).

Table 4.7. Restriction endonuclease cleavage data and associated Southern hybridisation data for plasmid construct pM3-3e (construct E)

Single - enzyme cleavage

Enzyme tested	No. of fragments	Sites in vector	Sites in insert	Fragment sizes (kb)	Sum of fragments (kb)	Identity* of fragment (site1-site2)	Hybridisation with probe 148-149
***Cla* I**	1	1	0				
***Eco*RV**	1	1	0				
***Kpn* I**	1	1	0				
***Sca* I**	0	0	0				
***Xho* I**	1	0	0				
***Bam* HI**	2	1	1	7.3		Bo - B1	
				3.5	10.8	B1 - Vo	
***Dra* I**	4	3	1	7.2		D - VD1	
		(2)		2.7		VD3 - D	
				0.7		VD1/2-VD3	
				0.2	10.8		
***Eco* R1**	3	1	2	6.2		E1 - E2	
				3.3		E2 - Vo	
				0.5	10.0	Eo - E1	
***Hin*d III**	5	1	4	7.2		H4 - Vo	-
				2.2		H2 - H3	++++
				1.1		Ho/1 - H1/2	-
				0.8		Ho/1 - H1/2	-
				0.5	11.8	H3 - H4	-
***Pst* I**	3	1	2	6.1		P2 - Vo	+
				3.0		P1 - P2	+++++
				2.4	11.5	Po - P1	-
***Sph* I**	4	0	4	4.2		S4 - S1	+/-
				3.2		S1 - S2	+++++
				3.2		S3 - S4	-
				1.3	11.9	S2 - S3	-
***Xba* I**	2	1	1	11.0		X1 - Vo	
				0.7	11.7	Xo - X1	

Average sum of fragments:	11.21	kb
***less* Size of vector:**	2.95	kb
Calculated size of cloned insert:	**8.26**	**kb**

* Cleavage sites are designated by the first letter of the relevant enzyme, followed by the order of the site(s) in the cloned insert, e.g. B_1 - V_0 represents the fragment spanning *Bam* H1 site (1) to the 3' end of the insert, which was cleaved at *Bam* H1 site (0, V_0) within the multiple cloning site of the vector. Similarly, H_2 - H_3 represents the fragment spanning *Hin*d III sites (2) and (3).

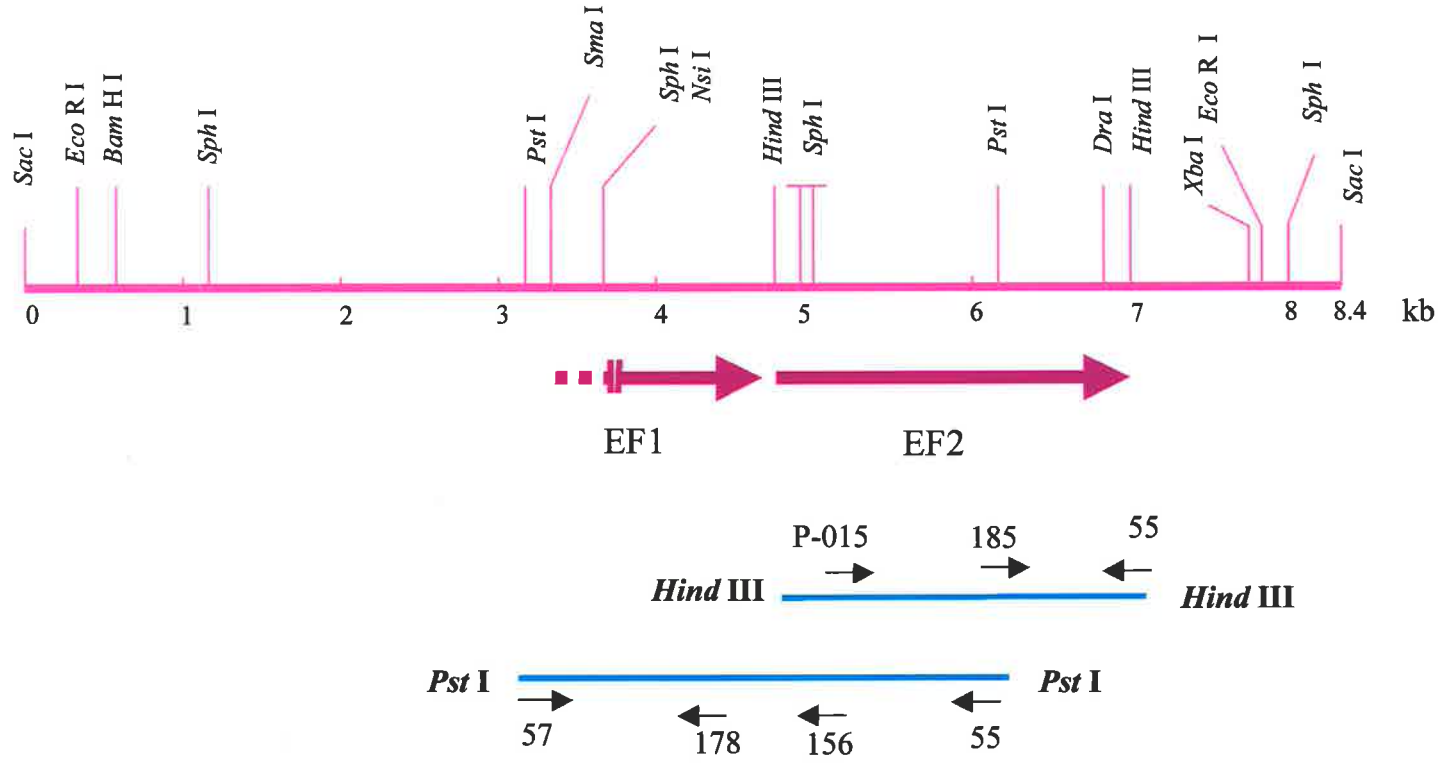


Figure 4.21. Schematic representation of the pM3-3E insert. Open reading frames EF1 and EF2 are denoted by thick arrows. The dotted line represents part of the EF1 reading frame which was not sequenced. Subcloned fragments used for sequencing are depicted by bold lines. Restriction sites that were identified experimentally by cleavage are shown. The identity and annealing position of primers (used for sequencing) are shown by short arrowheads.

Figure 4.22. Nucleotide and deduced amino acid sequences specified by open reading frames within the 2,792-bp insert of pM3-3E (construct E). The two open reading frames EF1 and EF2 are indicated, together with encoded C-terminal invariant segment 'CRGKA' and stop codons (both underlined), polyadenylation signal sequence (AGTRAA, double underlined) and the VSP-specific pre-polyA signal sequence, (CTTAGRT, underlined). This sequence was submitted to the GeneBank™ database with accession number AF236020.

```
EF1      N  Q  I  T  G  V  A  N  C  V  S  C  A  P  P  A  G  G  N  G  G  P
   1  ACAACCAAATCACTGGTGTCGCTAATTGCGTAAGCTGCGCCCCGCCAGCTGGTGGCAATGGCGGTCC
      V  T  C  Y  I  K  T  D  G  D  N  T  G  G  S  V  N  K  S  G  L  S  T
  68  TGTCACCTGCTATATCAAAACAGACGGCGACAACACCGGCGGAAGCGTCAACAAGAGCGGCCTCAGCAC
      G  A  I  A  G  I  S  V  A  V  V  V  V  V  G  G  L  V  G  F  L  C  W
 137  AGGTGCCATCGCGGGGATCTCCGTGGCTGTGGTCGTCGTCGTGGGGGGCCTCGTAGGGTTCCTCTGCTG
      W  F  I  C  R  G  K  A  *
 206  GTGGTTCATCTGTAGAGGGAAGGCGTGACTTAGGTAGTGAATGCTGTACAGTGATAGTGATAGTGCTAT
                                                M  G  G  C  Y  K  V  G  Q
EF2
 275  GTGTGGTGTGGGGGTGATAGTGATGGTGTGGGACTCCTCTTCATGGGCGGGTGCTACAAGGTAGGACAA
  10  D  P  G  N  E  I  C  T  A  A  D  K  G  V  C  T  T  C  K  A  G  A  A
 344  GATCCTGGCAACGAGATATGCACGGCGGCCGACAAGGGAGTGTGCACCACCTGCAAGGCAGGCGCCGCG
  33  Y  L  F  Q  N  P  A  S  P  V  A  P  G  N  E  C  I  L  C  S  D  T  T
 413  TATCTATTCCAGAACCCGGCCAGCCCAGTAGCCCCTGGCAACGAGTGCATCCTCTGCTCTGATACTACA
  56  Q  R  D  G  V  M  G  V  D  N  C  A  Q  C  Q  A  P  Q  S  A  G  A  A
 482  CAGAGAGACGGGGTCATGGGAGTGGATAACTGCGCACAGTGTCAGGCACCACAGAGTGCAGGAGCAGCT
  79  T  C  S  T  C  Q  D  G  Y  F  L  S  D  K  L  C  K  P  C  N  Q  N  C
 551  ACATGTAGTACCTGCCAAGATGGATACTTCCTGAGCGATAAGCTTTGCAAGCCATGCAATCAGAACTGT
 102  A  T  C  T  G  A  G  A  T  N  C  E  T  C  K  P  G  T  Y  L  K  S  D
 620  GCCACATGTACCGGGGCCGGTGCGACCAACTGTGAGACATGCAAGCCGGGGACCTATCTGAAGTCAGAT
 125  N  S  C  S  N  T  C  E  N  N  Q  Y  A  D  E  L  T  M  T  C  K  A  C
 689  AACTCATGCTCTAATACGTGTGAGAACAATCAGTATGCAGACGAGCTAACAATGACCTGCAAAGCATGC
 148  S  E  I  H  A  D  C  T  A  C  S  F  D  K  T  T  G  K  P  K  C  T  N
 758  AGTGAGATACACGCAGATTGCACGGCATGCTCCTTCGACAAGACCACGGGGAAGCCGAAGTGCACTAAC
 171  C  G  T  N  K  T  P  R  T  A  L  D  G  T  S  T  C  V  D  K  T  L  D
 827  TGCGGGACCAATAAGACCCCCAGGACGGCCCTCGACGGGACATCGACCTGCGTCGATAAAACGCTTGAC
 194  G  C  K  G  A  D  G  A  L  F  M  K  E  D  K  T  C  A  L  C  G  D  A
 896  GGGTGCAAGGGCGCTGACGGGGCCTTATTTATGAAGGAGGACAAAACGTGTGCTCTTTGCGGCGATGCG
 217  S  P  D  A  G  V  N  D  K  G  I  A  G  C  S  I  C  E  K  T  A  G  N
 965  TCCCCTGATGCTGGCGTAAATGATAAGGGCATTGCCGGGTGCAGCATATGTGAGAAGACGGCAGGAAAT
 240  P  P  T  C  S  K  C  L  E  G  Y  I  E  N  T  N  G  G  G  F  A  C  D
1034  CCGCCAACCTGTTCAAAATGCCTGGAGGGATACATCGAGAATACTAACGGTGGGGGTTTTGCGTGCGAT
 263  P  C  A  P  G  C  A  T  C  S  K  K  E  D  P  S  K  C  L  T  C  K  Q
1103  CCGTGCGCTCCAGGCTGTGCGACATGCTCCAAGAAGGAAGATCCGAGTAAGTGCCTGACATGTAAGCAA
 286  G  Y  F  L  K  D  S  S  S  G  E  C  I  S  C  I  D  T  T  K  V  A  S
1172  GGGTACTTCTTGAAGGATAGTTCTTCTGGTGAGTGCATCTCATGTATTGACACAACCAAAGTGGCATCG
 309  R  G  C  A  E  C  T  N  S  G  T  F  K  C  T  K  C  K  V  N  Y  R  P
1241  AGGGGGTGTGCCGAGTGCACAAATAGTGGCACATTTAAGTGCACAAAATGCAAGGTGAACTACAGACCC
 332  S  G  E  P  S  T  G  V  T  C  T  K  V  C  E  D  P  T  A  C  G  G  T
1310  AGTGGAGAACCTTCAACTGGGGTCACGTGTACAAAGGTGTGTGAAGACCCCACGGCCTGTGGTGGGACA
 355  S  G  A  C  D  A  I  V  I  D  N  T  G  K  E  L  H  Y  C  S  Y  C  G
1379  TCTGGAGCCTGTGATGCAATAGTAATCGACAACACAGGGAAGGAGCTTCACTACTGTTCTTACTGTGGA
 378  K  D  S  E  F  P  I  D  G  X  C  A  S  E  A  K  G  N  T  G  C  V  N
1448  AAGGACAGTGAGTTCCCTATTGATGGTNTCTGTGCTAGTGAGGCTAAAGGCAATACAGGATGTGTCAAT
 401  N  V  C  T  S  C  T  M  G  Y  F  L  Y  M  G  G  C  Y  S  I  S  A  Q
1517  AATGTCTGCACATCATGTACTATGGGATACTTCCTATACATGGGTGGTTGTTATAGTATATCAGCTCAG
 424  P  G  K  S  M  C  T  K  A  G  D  G  V  C  T  E  A  A  A  G  Y  F  I
1586  CCGGGTAAGTCCATGTGCACAAAGGCAGGTGATGGTGTCTGCACAGAGGCGGCAGCTGGGTACTTCATC
 447  P  P  S  P  T  K  D  K  Q  S  V  L  S  C  G  N  P  L  G  V  E  L  A
1655  CCTCCCTCGCCCACCAAGGACAAGCAGTCTGTCCTCTCCTGTGGGAACCCCCTTGGTGTCGAGCTGGCT
 470  G  Q  K  A  Y  V  G  V  D  G  C  S  Q  C  T  A  P  T  A  P  S  D  A
1724  GGTCAGAAGGCATACGTGGGGGTGGATGGCTGTTCACAGTGTACAGCCCCAACAGCTCCATCTGATGCT
 493  G  M  T  P  A  V  C  T  S  C  D  S  D  R  K  P  N  K  D  G  S  G  C
```

```
1793  GGCATGACCCCCGCTGTGTGTACCTCCTGTGACAGTGATAGGAAGCCCAACAAGGACGGCAGTGGGTGT
 516  V  T  C  S  V  G  D  C  K  S  C  V  M  D  D  V  C  G  E  C  S  D  G
1862  GTGACATGCTCTGTGGGTGATTGCAAGAGCTGTGTGATGGATGATGTGTGTGGGGAGTGCAGTGATGGG
 539  Y  S  L  E  G  G  K  C  V  S  S  G  A  N  R  S  G  L  S  T  G  A  I
1931  TACAGTCTGGAGGGTGGTAAGTGTGTGTCCTCTGGCGCCAACAGGAGCGGCCTCAGCACAGGCGCCATC
 562  A  G  I  S  V  A  V  V  V  V  V  G  G  L  V  G  F  L  C  W  W  F  L
2000  GCGGGGATCTCCGTCGCTGTGGTCGTCGTCGTGGGGGGCCTCGTGGGCTTCCTCTGCTGGTGGTTCCTC
 585  C  R  G  K  A  *
2069  TGCAGGGGGAAGGCGTAGACTTAGCTGCGTACTTAGGTAGTAAATGCTGTGTGGTGGACAGGTGAGCAG
2138  TGTTCTGGCGTGGGCGAACAGTGGGTGCCTAACGCCTCACACTGGTGGGCCGGGGCGCTGACCCAGGGT
2207  CCTCCCAGCCGGGCGGCGCGACTCGTGCTGGCGTGCAACCATGGAGCTGCGGACTAAGAAGACAAGGAC
2276  TGCATGGTGGAGAGACCGCAGATAAGCGCAGTCCACATCCCTACATTGCTGTAGAGCAATGCACCGGTG
2345  TCAAAGGTAGGGGGTTGTCTAGGCTGATAGCTTTATAGTGCAAAGCAAAGGATCAAGCGGAAAAGACCT
2414  ATAAATGGTTGGACAAATCGTTGTCACCCATAAAGAAAGGCATCTCCAATGACCCCACGATCTGCAAAG
2483  ACTGTCTGATGATTTACAGTGGCATTTGCCTCCTTGTCAGTTGAAACAGACACAAGTTAAGACGGAGTA
2552  AATCTCTAATTTAAAAACTGATCTCCAGTTATCTTCTAACACGGGTGTGCCGTGTATGGATCAAATTCG
2621  CATCACGAACCCAAGCCAATTCAGAGCAGCACTTCAAGCACTTNTAATCGCTGACCAGCCCAATATCTG
2690  GGTTTCATTGCGACAGCAAGATAATTTGACGCTTACATATATGGAGTCGGATAGAGATGGCTTTGAAGG
2759  GCTTGCAGTTCGAATGGCAGAAGATGTATGTATT
```

poly(A) signal sequence). The EF2 reading frame ended with a TAA codon and two additional in-frame stop codons (both TAG, the last part of the extended poly(A) signal sequence) were located 3 bp and 15 bp further downstream respectively (**Fig. 4.22**).

The EF2 sequence, although commencing with a putative methionine start codon, lacked a functional 5' end, i.e. it did not encode a hydrophobic signal peptide and on this basis the ORF appeared to be a pseudo gene. Examination of the sequence upstream from the start of the EF2 reading frame (including the 3' end of EF1) failed to reveal any overlapping reading frame that might have encoded the missing N-terminal segment (data not shown). Moreover, numerous stop codons were identified in all three reading frames within the 100-bp segment immediately upstream from the EF2 initial methionine codon. These included the three (mentioned above) at the 3' end of the EF1 reading frame; six in the –1 (EF2) frame (four adjacent codons [TGATAGTGATAG] located 16 codons [48 bp] upstream from the start of the EF2 ORF, a fifth located 36 codons upstream [between the EF1 stop codon and its polyadenylation signal] and the sixth situated a further 7 codons upstream, overlapping the last EF1 Arg (-CRGKA) codon); and four in the –2 frame, including three adjacent codons (TGATAGTGA) 19 bp upstream from the the start of the EF2 ORF (**Fig. 4.22**). These findings indicate that the EF2 gene lost functionality by recombinational deletion of its 5' end, not by frameshift mutations. The sequence of the 3' segment of the EF1 reading frame showed no close similarity to EF2 or the M165-[3'] probe sequence. This was consistent with Southern hybridisation data on pM3-3E restriction fragments, which revealed that only the 1.95-kb *Hin*d III fragment hybridised with the probe (**Tables 4.7 & 4.8**).

4.9 Analysis of other cloned fragments

It was beyond the feasibility of this Ph.D. project to attempt a comprehensive analysis of all the 20 cloned *Sac* I genomic inserts identified and described earlier (**Table 4.2**) as containing sequences related to the pM165 insert. Indeed, to pursue this ambitious goal

would need to have been justified by very solid strategic considerations. Nevertheless, because these additional uncharacterised inserts were identified by colony blotting and therefore known to contain sequences that were specifically related to the pM165-[5'] and pM165-[3'] probes, experiments were undertaken to obtain more general information about the identity of the respective inserts. All of the 13 remaining constructs had been examined initially at a superficial level by restriction analyses which had shown them to be unique, as indicated by the summarized data in **Table 4.2**. Several inserts were also examined by Southern hybridisation analysis of restriction digests - as exemplified in **Fig. 4.23** for pM17-9, which contained a 6.6-kb insert. However, none of these findings provided any indication of the nature or number of vsp gene (or pseudo gene) sequences that were present in these cloned *Sac* I fragments.

With the aim of identifying specific vsp gene sequences within these remaining 13 fragments during the final 2-3 months of the project, a new approach was adopted. The strategy taken for this broader, less detailed survey was to:

1. Screen the *Giardia* genome database (comprising random sequences considered to represent approximately 30% of the genome) for sequences that were related to *crp*72 or to the pseudo genes identified in this study.

2. Align the deduced amino acid sequences to identify specific or distinct differences between the N-terminal segments encoded by putative functional genes and those specified by the pseudo genes.

3. Use these differences (or homologies) to design PCR primers for use in identifying particular functional or pseudo genes, or to distinguish the two types of coding sequences.

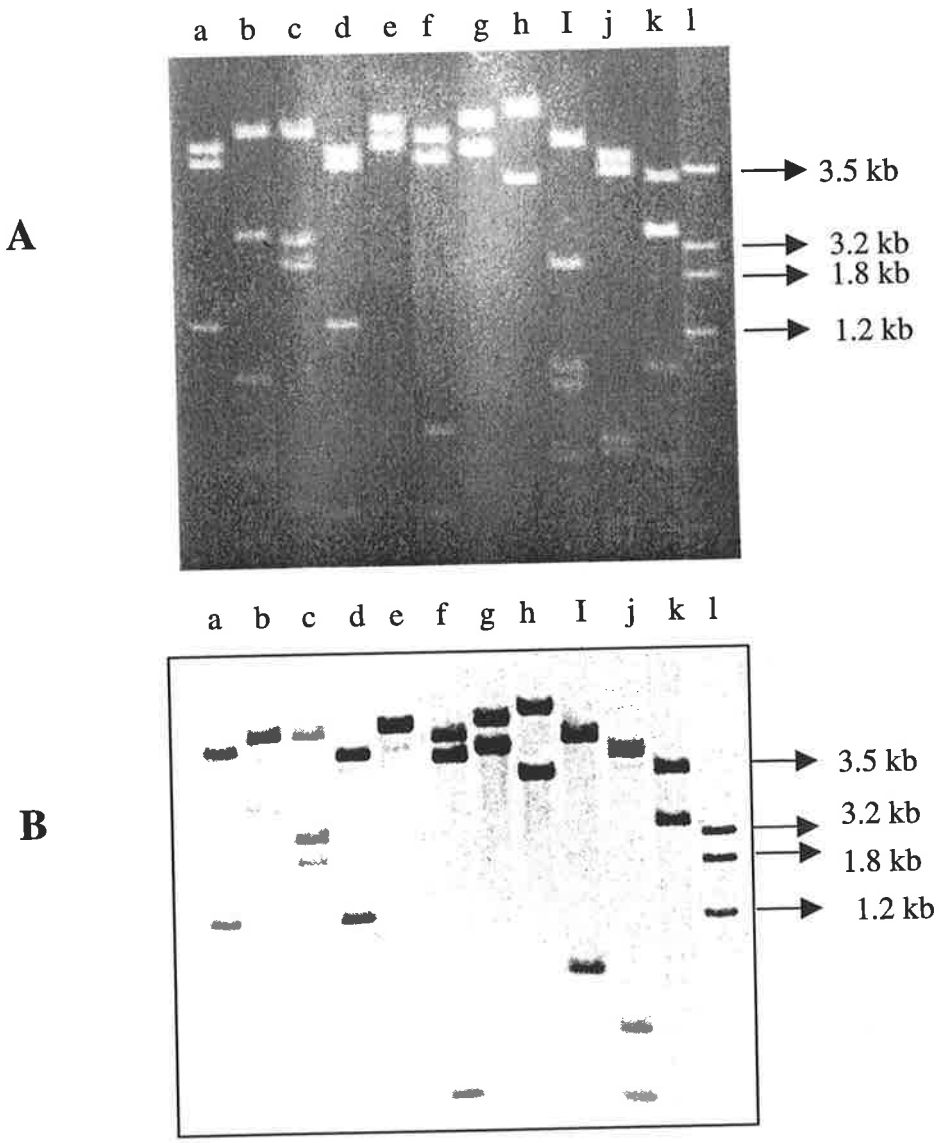


Figure 4.23. Restriction and Southern hybridisation analysis of pM17-9. The construct contained a 6.6-kb *Sac* I insert derived from *Giardia* genomic DNA. Replicate aliquots of plasmid DNA were incubated with different restriction endonucleases. The reaction mixtures were subjected to electrophoresis on 1% agarose and stained with ethidium bromide (A). The gel was subsequently analysed by Southern hybridisation (B) using the DIG-labelled M165-[5'] probe. Size markers are indicated. The enzymes used were :

a	*Bam* H I	h	*Xba* I
b	*Cla* I + *Dra* I	i	*Bam*H I + *Dra* I
c	*Kpn* I	j	*Sph* I + *Cla* I
d	*Pst* I	k	*Dra* I + *Xba* I
e	*Sma* I	l	*Xba* I + *Pst* I
f	*Sph* I		
g	*Xho* I		

4. Test the various constructs in PCR using these different primer combinations to determine which coding sequences (or vsp gene types) might be present within the cloned inserts.

The rationale behind this approach was the identification, in the genome database, of particular sequences from closely-related functional vsp genes. Several related sequences were found in the database. Alignment of these retrieved sequences with the *crp*72-like pseudo gene sequences discovered and characterised in this project revealed conserved and variable segments. These were exploited to design subset-specific primers. Some of these oligonucleotides were designed to anneal only to apparently functional *crp*72-related genes, whilst others were designed to hybridise to both functional and "5'-defective" pseudo genes e.g. **Fig. 4.24**. Using BLAST, the *Giardia* genome database was searched for sequences similar to the first 400 bp of *crp*72. Six DNA sequences were found, each exhibiting 96-99% nucleotide identity with this 5' segment of the *crp*72 coding sequence.

One of these database sequences, KI0466SA, was identical at 99.6% of the first (5') 785 bp of *crp*72. The *crp*72 transcript sequence reported by Adam *et al.* (1992) remains incomplete, lacking the first few codons of the gene. A comparison of the KI0466SA and *crp*72 nucleotide sequences indicated that if, as appeared likely, the KI0466SA sequence was derived from the *crp*72 locus, the first 35 nucleotides that were missing from the 5' end of the published *crp*72 sequence would be: ATG TTC CAA CTG ATA CCC CTG TTC GTA GCG AGC GC.

On the basis of multiple alignments of both the nucleotide and inferred amino acid sequences derived from the characterised ORF of the *Sac* I genomic inserts, the provisionally complete *crp*72 and five additional related sequences retrieved from the genome database, several primers were synthesised (**Fig. 4.25**). Primers 199, 193 and 194 were designed to amplify only the extreme 5' segments of *crp*72 or closely related *functional* vsp genes. These segments (inclusive of the primer sequences) are missing from the pseudo genes described in

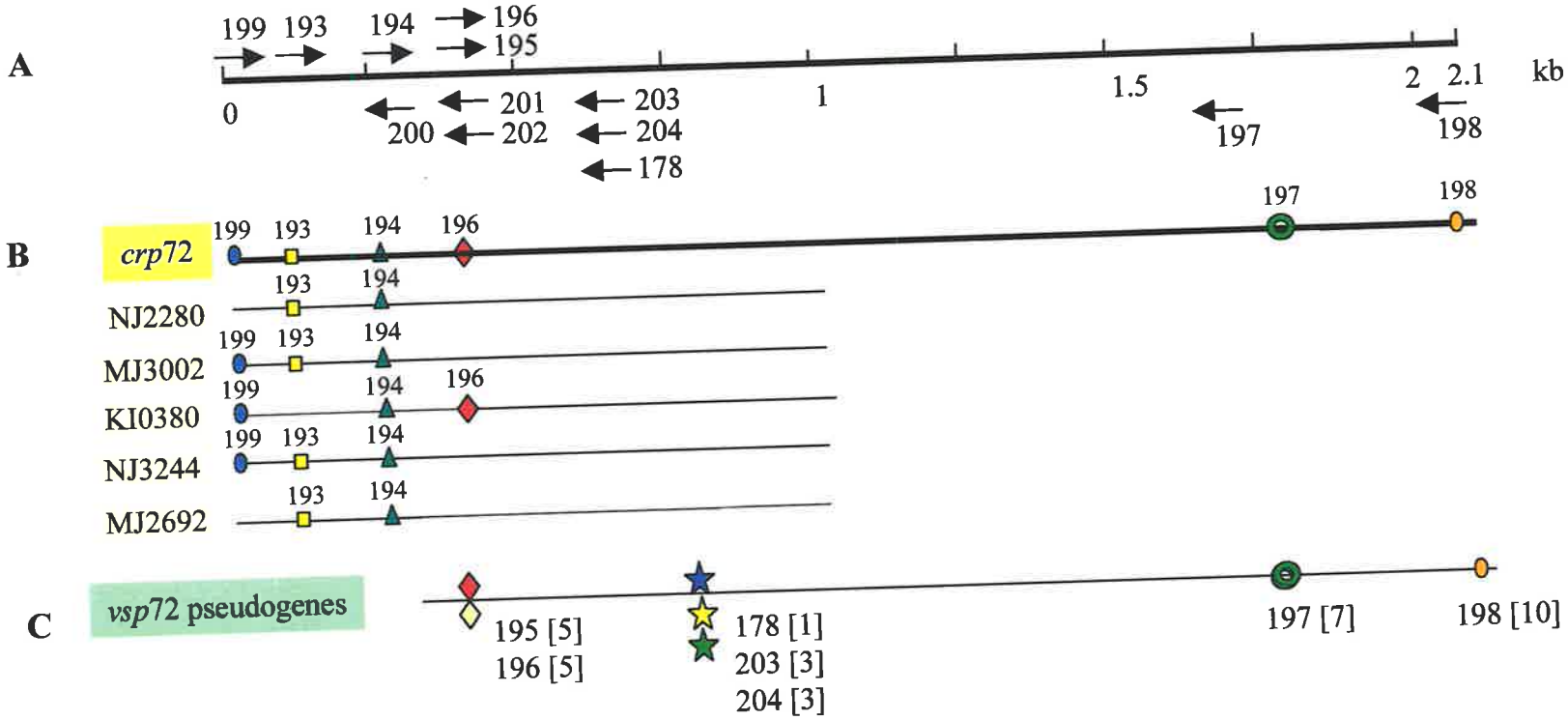


Figure 4.24. Scale diagram depicting the locations of primer sites within the corresponding 5' segments of *crp*72-like gene sequences (derived from the genome database) and related pseudogene sequences identified in this study. Approximate annealing locations of PCR primers (arrows) are indicated on a linear scale (A), corresponding to the 2.1-kb *crp*72 coding sequence, or as geometric symbols on putative functional (B) or 5'-defective pseudogene (C) sequences. The absence of a symbol in B or C indicates the absence (mismatch) of the primer sequence from that particular genes.

the previous section (**Figs. 4.24, 4.25**). Other primers were designed on the basis of downstream consensus sequences to amplify segments from these latter *crp*72-like pseudo genes as well as from the corresponding portions of their functional counterparts. Figure 4.24 illustrates the relationships between these primers, the location of sites to which they hybridise in the pseudo- and functional genes, and the locations of downstream reverse primer sites. These primers were used to examine the remaining 13 constructs, which were not subjected to nucleotide sequence determinations, for the presence or absence of specific gene segments.

The results of these experiments are summarised in **Table 4.9**. On the basis of the pseudo genes described herein, three primers (199, 193 and 194; **Fig. 4.25**) were expected to amplify only functional (intact) *crp*72-like genes. As can be seen in **Fig. 4.24**, the putative functional *vsp*72-like sequences that were identified in the *Giardia* genome database contain one or more of these primer sequences. This was used as a diagnostic guide to interpret the PCR results. The combination of primer 199 and 197 was used to test for the amplification and identification of a ≈1.65-kb segment of presumed functional vsp genes belonging to the *vsp*72 subfamily (**Fig. 4.24B**). It was hoped that this might provide a preliminary indication of how many of the cloned inserts contained such functional *vsp*72-like genes. It should be noted, however, that, not all of the identified (putative) functional *vsp*72-like gene segments encoded the same N-terminal (leader signal peptide) sequences. Hence, whilst the recovery of a '199-197' amplification product might reasonably be considered to indicate that a given functional or pseudo gene is present within a particular construct, any *failure* to amplify this segment does not necessarily mean that a particular insert contains no functional *vsp*72-like gene. **Table 4.9** shows that only 4 of 13 inserts tested with primers 199 + 197 yielded a PCR product in the expected size range (≈1.6 kb). Using primers 193 or 194, each in combination with reverse primer 197, two and eight constructs respectively (from the panel of 13) yielded PCR products in the expected (1.5-kb) size range. PCR amplification using forward primers

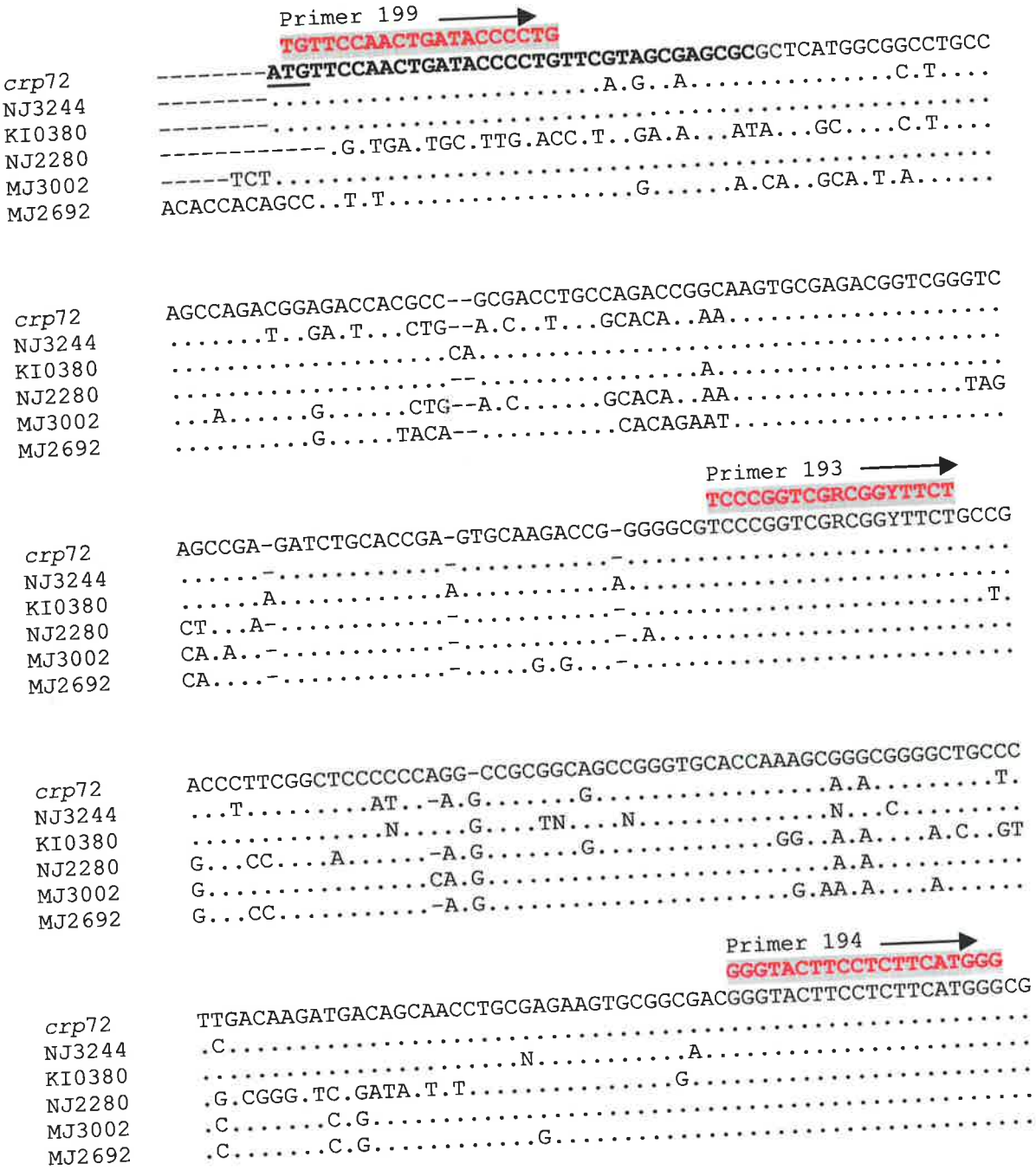


Figure 4.25. Alignment of *crp*72-like VSP gene 5' Nucleotide sequences retrieved by BLAST searches of the *Giardia* genome database. Default settings of CLUSTAL W were used and the output was corrected by manual editing. Gaps (-), were introduced to maximise the alignment. Dots indicate nucleotide identity between *crp*72 (*vsp*1269, row 1) and corresponding positions in the other sequences. The presumptive first 35 nucleotides of *crp*72 were obtained from the Genome database (sequence ID: KI0466SA) and these are shown in bold type. The putative initiation codon of *crp*72 is underlined.

Table 4.9. Summarised PCR data indicating the presence or absence of functional or 5'-defective pseudogenes in cloned *Sac* Igenomic restriction fragments.

Plasmid construct[a]	Primers used for PCR[b]					Functional genes	Multiple VSP cds
	199+197	193+197	194+197	195+197	196+197		
5-1	-	-	+	+	+	likely	+
17-9	-	+	-	+	+	likely	+
19-2	-	-	-	-	+	unlikely	?
3-2	-	-	+	+	+	likely	+
14-3	-	+	-	-	+	likely	?
17-1	+	-	+	+	+	yes	+
16-3	-	-	+	-	+	likely	?
17-4	-	-	+	+	+	likely	+
17-7	+	-	+	+	+	yes	+
12-8	-	-	-	+	+	unlikely	+
21-1	+	-	+	+	+	yes	+
16-1	+	-	+	+	+	yes	+
9-2	-	-	-	+	+	unlikely	+

[a] These contained inserts which had been characterised incompletely restriction mapping and also (for some inserts only) by Southern hybridisations.

[b] Tested for amplification of specific *crp* 72-like gene segments using different primer combinations as indicated (see Fig. 4.24). Recovery of (or failure to recover) a product of the expected size is denoted by (+) or (-) respectively.

195 and 196 (which should anneal to conserved sequences present in both functional and pseudo genes) in combination with primer 197, the expected 1.3-kb PCR products were amplified from 10 (primer 195) and 13 (primer196) of the 13 cloned inserts. In view of the fact that DNA of the expected size was amplified in high yield from 10 of the 13 constructs tested in PCR using primers 195+197 or 196+197, it appeared likely that most of the inserts possessed more than one *vsp*72-like gene sequence (**Table 4.9**).

4.10 Identification and characterisation a functional *vsp*72-like locus in pM21-1 (construct Q)

As mentioned above (section 4.9), only four of the 13 constructs tested in PCR using primers 199 + 197 yielded a product similar in size to the 1.6-kb segment expected for functional *vsp*72-like genes (**Table 4.9**). Although these results were obtained in the final stages of the project when time was very limited, it was deemed important (in view of the pseudo gene arrays detected by nt sequence analysis in five cloned *Sac* I fragments, as described in section 4.8) to confirm the existence of a functional *vsp*72-like gene within one of these four genomic fragments. Plasmid construct pM21-1 (**Table 4.9**) was chosen for this purpose. The 1.64-kb DNA that was amplified in PCR from this template DNA using primers 199+197 (**Fig. 4.24**) was subjected to sequence analysis by primer 'walking'. The compiled 1,644-bp sequence, shown in **Fig. 4.26**, represents an uninterrupted reading frame that appears to be part of a longer sequence. It is notable that this elucidated sequence extends 246 bp upstream from the 5' truncated EF2 pseudo gene identified in pM3-3E (section 4.8.2e), i.e. the N-terminal amino acid sequence 'MGGCYK-' specified by the EF2 pseudo gene corresponds to residue 83 onward of the polypeptide ('PQ1') encoded by the partial pM21-1 sequence (c.f. **Figs. 4.22 & 4.26**). Because this latter sequence was determined from an amplified internal segment of the genomic fragment cloned within the pM21-1 construct,

Figure 4.26. Nucleotide and deduced amino acid sequences corresponding to part of a putative functional *crp*72-like gene. This member of the vsp72 gene subfamily was identified within the 1,644-bp insert of pM21-1 (construct Q). This sequence was submitted to GeneBank™ database with accession number AF298862.

```
   1  F  Q  L  I  P  L  F  V  V  S  A  L  A  V  T  C  Q  A  D  K  C  E  T
   1  TTTCAACTGATACCCCTGTTCGTGGTGAGCGCGCTTGCAGTAACCTGCCAGGCCGATAAGTGCGAGACG
  24  V  G  N  T  E  I  C  T  Q  C  R  A  R  G  V  P  V  D  G  F  C  W  P
  70  GTCGGTAACACAGAGATCTGCACCCAGTGCAGGGCCCGGGGCGTCCCGGTCGACGGCTTCTGCTGGCCC
  47  P  G  F  P  Q  A  A  A  A  G  C  T  E  E  D  G  V  P  L  D  K  T  A
 139  CCCGGCTTCCCCCAGGCCGCGGCAGCCGGGTGCACCGAGGAAGACGGGGTACCCCTCGACAAGACGGCA
  70  A  T  C  G  K  C  G  D  G  Y  L  L  F  M  G  G  C  Y  K  T  E  S  Q
 208  GCCACCTGCGGGAAGTGCGGCGACGGGTACCTCCTCTTCATGGGCGGATGCTACAAGACAGAGAGTCAG
  93  P  G  S  D  I  C  T  A  A  S  N  G  V  C  T  E  C  N  T  K  N  G  L
 277  CCCGGCAGTGACATATGCACAGCAGCCAGTAACGGAGTATGCACTGAGTGTAATACAAAGAACGGCCTG
 116  F  K  N  P  A  T  A  P  E  K  G  R  E  C  I  L  C  H  D  A  T  G  A
 346  TTCAAGAACCCTGCTACAGCACCAGAGAAAGGAAGGGAGTGCATACTCTGTCATGATGCCACGGGAGCA
 139  D  G  Y  M  G  V  E  G  C  A  T  C  T  A  P  T  N  N  K  G  A  A  T
 415  GATGGATATATGGGAGTCGAGGGCTGCGCTACATGTACAGCACCAACTAATAATAAGGGGGCAGCCACG
 162  C  T  E  C  Q  D  G  Y  Y  N  D  G  G  A  C  K  K  C  V  D  G  C  I
 484  TGCACAGAGTGTCAAGATGGGTACTACAACGATGGTGGCGCATGTAAGAAGTGCGTTGATGGATGCATA
 185  D  C  T  G  A  N  Q  C  T  T  C  E  D  G  K  Y  L  K  N  N  Q  C  V
 553  GACTGCACCGGTGCAAATCAATGCACAACATGTGAGGACGGGAAATACCTGAAGAATAATCAATGTGTA
 208  D  A  G  Q  C  D  Q  G  T  Y  A  D  P  T  T  G  Q  C  K  P  C  G  I
 622  GATGCTGGTCAATGTGATCAAGGCACCTATGCAGACCCGACAACAGGCCAATGCAAGCCATGTGGAATA
 231  T  D  C  A  T  C  E  Y  N  A  T  I  S  Q  P  Q  C  K  T  C  S  T  S
 691  ACTGACTGTGCCACCTGTGAGTACAATGCAACTATTAGTCAACCACAGTGTAAGACCTGCAGTACTAGT
 254  S  N  K  M  V  K  T  A  A  D  G  T  T  T  C  V  D  D  G  G  C  T  N
 760  AGTAACAAGATGGTGAAGACAGCGGCAGACGGGACGACCACTTGTGTCGATGATGGTGGATGTACAAAT
 277  G  N  T  H  F  V  E  G  T  N  Q  K  L  C  V  P  C  G  D  T  T  N  G
 829  GGCAATACGCACTTTGTTGAAGGTACTAACCAAAAGCTTTGTGTCCCATGTGGTGATACTACAAATGGT
 300  G  V  L  G  C  N  T  C  S  S  K  T  T  C  T  K  C  L  D  G  Y  Y  D
 898  GGGGTTCTAGGTTGTAATACTTGCTCCTCTAAAACTACATGTACAAAGTGCCTCGACGGATACTACGAT
 323  S  G  S  G  T  V  T  C  T  A  C  P  G  A  N  C  A  T  L  C  E  R  Y
 967  AGTGGTAGTGGTACGGTTACATGCACCGCTTGTCCAGGTGCTAACTGTGCCACCCTGTGTGAAAGATAC
 346  K  R  Q  C  T  T  C  K  P  G  F  F  L  K  D  S  S  S  G  E  C  I  S
1036  AAAAGACAATGCACGACCTGTAAGCCTGGGTTCTTCTTGAAGGATAGCTCCTCTGGTGAGTGCATCTCA
 369  C  S  D  K  N  N  G  G  H  E  G  C  S  A  C  S  S  N  G  A  F  K  C
1105  TGTAGTGACAAAAACAATGGGGGCCACGAAGGGTGCAGTGCATGCTCTAGCAACGGTGCTTTCAAGTGC
 392  T  D  C  K  P  N  Y  K  K  E  G  T  S  D  N  Y  T  C  V  K  T  C  E
1174  ACTGACTGTAAGCCTAACTACAAAAAAGAGGGAACTTCAGACAACTATACATGTGTGAAGACTTGTGAG
 415  D  E  T  A  C  G  G  T  S  G  A  C  D  A  I  V  I  D  E  N  G  N  T
1243  GACGAGACGGCTTGTGGTGGTACTTCTGGGGCCTGTGATGCAATAGTAATTGACGAGAATGGCAACACA
 438  K  H  Y  C  S  F  C  G  E  S  G  K  F  P  I  D  G  L  C  A  S  D  K
1312  AAGCATTATTGTTCATTCTGTGGGGAGAGTGGTAAGTTCCCTATAGATGGTCTCTGTGCTAGTGACAAG
 461  A  N  N  N  G  C  A  N  G  V  C  T  S  C  T  A  A  N  Y  F  L  Y  M
1381  GCTAATAATAATGGATGTGCCAATGGTGTCTGTACATCATGTACTGCTGCTAACTACTTCCTCTATATG
 484  G  G  C  Y  K  V  N  T  V  P  G  S  H  M  C  K  T  A  N  N  G  V  C
1450  GGTGGTTGTTATAAGGTTAATACAGTACCAGGTAGTCATATGTGCAAGACAGCTAACAACGGTGTCTGT
 507  T  A  V  S  E  N  N  K  Y  F  I  V  P  G  A  S  N  Q  N  Q  S  V  L
1519  ACAGCTGTTAGTGAGAACAATAAGTACTTCATAGTCCCAGGGGCATCCAATCAGAATCAGTCTGTACTA
 530  A  C  G  N  P  L  G  T  E  L  T  A  K  A  Y  V  G  V  K
1588  GCCTGTGGTAACCCCCTAGGCACTGAATTAACTGCTAAGGCATACGTTGGGGTTAAA
```

the ORF evident in **Fig. 4.26** can be expected to extend further upstream. The probability that this ORF (designated pQ1) represents a functional vsp gene is considered in the next section.

4.11 Comparison and phylogenetic analysis of the characterised loci of the '*vsp*72' gene subfamily

To further examine the relationships between the identified *vsp*72-like loci, a sequence based phylogenetic analysis was undertaken. Initially, the amino acid sequences encoded by the available *vsp*72-like loci were aligned using CLUSTAL W. The alignment was edited manually to include the N-terminal segments of CRP72 and PQ1 (section 4.10) and to highlight the multiple stop codons that were found to precede many of the vsp pseudo gene sequence. The resulting multiple sequence alignment is presented in **Fig. 4.27**. This shows the aligned putative amino acid sequences of CRP72 and the VSP72-like polypeptides specified by the 14 open reading frames that were identified and characterised in this project.

The alignment highlights the lack, from all 13 putative pseudo genes, of a functional 5' end that includes a valid initiation codon and the segment encoding the first 68 (for psC2) to 105 amino acid residues (for psA2 and at least 7 other loci) of the N-terminal segment of related functional genes. This is evident by the comparison with CRP72 (**Fig. 4.27**). Two pseudo genes, psC2 and psE2, appear to extend slightly further (20-30 codons) upstream than the others, most of which specify polypeptides commencing with the conserved amino acid sequence '(W)SEICT-'. Inspections of all three coding frames at the 5' ends of those pseudo genes for which sufficient sequence data was available failed to identify any inferred amino acid motifs that resembled the segments that were missing, based on comparisons with the corresponding region of CRP72 (**Fig. 4.27**). In addition, for several of these pseudo genes (those identified within a gene array) the 3' end of an adjacent (upstream) gene lay within 13-16 bp of the 5' 'start' of the pseudo gene (discussed in earlier sections of this chapter). In these cases, the missing segments would have to overlap the 3' ends of the adjacent genes.

Figure 4.27. Amino acid sequence alignment of CRP72 and the CRP72-like polypeptides specified by the VSP gene sequences identified from cloned *Sac* I fragments of genomic DNA from the Ad-1/c3 isolate of *G. intestinalis*. Gaps (-) were introduced to optimize the alignment. Dots indicate amino acid identity between CRP72 (top row) and corresponding positions in the other deduced polypeptides. Stop codons that precede the pseudo (ps) gene ORFs are depicted by asterisks (*). The C-terminal hydrophobic (presumptive membrane-spanning) segment is shaded grey. The predicted N-terminal signal (leader) peptide sequence of CRP72 and the invariant segment '-CRGKA' at the C-terminal ends of the polypeptides are underlined.

```
CRP72   MFQLIPLFVASALMAACQPDGDHAATCQTGKCETVGSAEICTECKTGGVPVDGFCRPFGS
pQ1     -........V...----------.V...AD......NT....Q.RAR........W.P.F
psB2    ----------------------------------------------------------
psA1    ----------------------------------------------------------
psC2    ----------------------------------------------------------
psE2    ----------------------------------------------------------
psA2    ----------------------------------------------------------
psA4    ----------------------------------------------------------
psA3    ----------------------------------------------------------
psC3    ----------------------------------------------------------
psD2    ----------------------------------------------------------
psD3    ----------------------------------------------------------
psC4    ----------------------------------------------------------
psB1    ----------------------------------------------------------
psD1    ----------------------------------------------------------

CRP72   PQAAAAGCTKAGGAALDKMTATCEKCGDGYFLFMGGCYKTTDGPGSEICTKAEGGLCTEC
pQ1     .........EED.VP...TA...G......L.........ESQ...D...A.SN.V....
psB2    ----------------------------------------**WV...A.SN....A.
psA1    ----------------------------------------------------------
psC2    -------*MLYSDSDS.SAMWCGGDN.GVGLF........ESQ...D...A.SN.V..T.
psE2    ----------*CYVWCGGDSDGVGL........VGQD..N....A.DK.V..T.
psA2    ----------------------------*---------**W.....A.SD....A.
psA4    ----------------------------*---------**W.....A.SD....A.
psA3    ----------------------------------**----**W.V...A.SD.V..A.
psC3    ----------------------------------------W.....Q.SD....A.
psD2    ------------------------**---------**----**W..........V....
psD3    ------------------------**---------**----**W..........V....
psC4    ----------------------------**---------**W.V...A.SD.V..A.
psB1    ----------------------------------------------------------
psD1    ----------------------------------------------------------

CRP72   KTANG-LFKNPAATPEKGSECILCSDINGADGYTGVANCAQCTKSDSNKGAATCTACQAG
pQ1     N.K.........TA....R.....H.AT.....M..EG..T..APTN........E..D.
psB2    ..D.KYV.Q...TA....R.....H.AT............T.................
psA1    -----.----------------------------------------------------
psC2    .AD.........VTLQKP.........TK............LK..EPTNSP......E..E.
psE2    .AGAAY..Q...SPVAP.N.......TTQR..VM..D.....QAPQ-SA.....ST..D.
psA2    ............TA....R.......TTDR..V..A.G.SE.SHTGTS-.P....V..D.
psA4    ............TA....R.......TTDR..V..A.G.SE.SHTGTS-.P....V..D.
psA3    N...QYI.Q.K.T.VTP.G.....H.AT..NENK.....LK..APANSP......E.MS.
psC3    .A..QYI.Q.R.ERVTP.G.....H.AT..NENK.....LK..APANSP......E.MS.
psD2    N.....................H..K..............QAPQ-SA.PV......D.
psD3    N.................R.....................ST.P...ST.RD.
psC4    N...QYI.Q.K.T.VTP.G.....H.AT..NENK.....LK..EPSGSP......E....
psB1    -----.----------------------------------------------------
psD1    -----.----------------------------------------------------

CRP72   YYKD-FQACSKCDGTCLTCETS-AAQCTSCPEGKYLKGDKSCVNNNGCTGNTYADPESGK
pQ1     ..N..GG..K..VDG.ID.TG-..N...T.ED.....NNQ-..DAGQ.DQG.....TT.Q
psB2    ..................................T..TSDQ..STV.P.....T
psA1    ----.----------------.-------------------------------------
psC2    ...E.NNE.NQ..QS....SG.GPNH....K......S.NT.SP--T.E.......VT--
psE2    .FLS.DKL.KP.NQN.A..TGAG.TN.ET.KP.T...S.N..S.--T.EN.Q...ELTMT
psA2    .I.K.GD..E...QS....DA.GPN............SNQ-..QDT..D..........
psA4    .I.K.GD..E...QS....DG.GPNH...........T.....................
psA3    FSG..-SCATQ.G.D.AA.DK.NQN.....KT.....ENQ-..EKSA.NN.H.P.DT.MT
psC3    FSG..-SCATQ.G.N.AA.DK.NQN.....KT.....ENQ-..EKSA.NN.H.P.DT.MT
psD2    FV.K.DS..V..GEG.SA.SADTPT...A.V...F..A..-..DA.Q.DNGK....KT.Q
psD3    ..M..S...T..NDN.A..TGAGQN..S..KA.F...S.G..SK--T.DN.Q.P..ST..
psC4    ....GNG..V..NEA....SGEG.TK..F.AGE...D.NN-..DASS.N.DK.PN.ST..
psB1    ----.----------------.-------------------------------------
psD1    ----.----------------.-------------------------------------
```

```
CRP72   CLPCNTID-----QACTQCEVDSTTKKPKCTNC--GGQKMVKTAIDGTTTCVDANGCATS
pQ1     .K..GIT.....--.AT..YNA.ISQ.Q.KT.STSSN.....A.....DG..TNG
psB2    .KE.S..........T.KYNA.VS..Q..A.NN--N.K.............I-SSGCE
psA1    --------....--------------------..--------------------SSNCV
psC2    -RT.KESG....IAD..A..YNT.VS..Q..A.NN---.K...EL.......DA....D
psE2    .KA.SE.H.....AD..A.SF.K..G.......G.-TN.TPR..L...S....KTLDGCK
psA2    .KA.S..H.....AE..A.SF.KG.G.......G.-AS.IPR.TL...S...TKGYTECQ
psA4    ............................A.DNS.-.K.............VS-----
psA3    .VA.S..-......D..A.TM.QS.G.......G.-AS.IPR.TL...S...TKGYTECQ
psC3    .VA.S..-......D..A.TM.QS.G.......G.-AS.IPR.TL...S...TKGYDQCQ
psD2    .KA.TDTS....VNE.AT.AYSD.LQ..V..G.NS..NLLL.VNP..SA...AEAE--CT
psD3    .TA..GAAEQGGIPE..A.TY.AKLQ..V.SD...-.N.K...EL.....S.HMS-----
psC4    .TA..AGADQGGIPE..A.TYNAKLQ..V.SA.NG---.K...EL........DA.--CK
psB1    --------....--------------------..--------------------SSGCV
psD1    --------....--------------------..--------------------SSGCV

CRP72   NVDGSHFLNDGSTK-CILCSDDSESL--EA-NKGTPGCKTCKKNG--AKPTCSECLDGYY
pQ1     .T---..VEGTNQ.L.VP.G.TTNGGVL---------..N..---SS......TK......
psB2    DA.--..KA..DAA..V....T.G.DP.--K...IA...A.T.TAS..N.......A...
psA1    DA.--..KA..DSA..V....T.G.DP.--K...IA...A.T.TAS..N.......A...
psC2    ........S.D..........TTTGT....ND..IAN.................A.....F
psE2    GA..AL.MKEDK-T..A..G.A.PDAGV--ND..IA..SI.E.TAG.NP....K..E..I
psA4    GA.KEL.MKEDQSA..L..G.TK.AS..--ND..VAN.R..T..ANDSP...TA..N..F
psA3    ------------.----------..--.-------------.------------
psC3    GA.KEL.MKEDQSA..L..G.NT.DTSESNK.Q...N....T.TAS.T..V.ET.K..FF
psD2    GT.KEL.MKEDQSA..L..G.NT.DTSESNK.Q...N....T.TAS.T..V.ET.K..FF
psD3    SGNT-...EQSPKA..VP.G.T----..--K.G.IL..D..S---..S.T..TK......
psC4    ------------.----------..--.-------------.------------
psB1    KDST-..VD.DN.M..V....T.GNEP.-NK...LK...A.T.QSS.SP...TG..P...
psD1    DEN--..KA.DDSA..Y..G.I.GNN...NT...VAD.NK.T.EAG.T.T......S.RF
        DEN--..KA.DDSA..Y..G.I.GNN...NT...VAD.NK.T.EAG.T.T......S.RF

CRP72   NSGNGGTVTCEACGANCATCTQAGN-DKCTKCKPGFFMKGNGPT--GECVACDNAQG-GI
pQ1     D..S-.....T..........ERYK-RQ..T......L.DSSS-.....IS.SDKNNG.H
psB2    S..S.-....A...D.......T...NQ...........S.....................
psA1    S..S.-....T...Q...........NQ.....A...L.....-......P..D..NQ..
psC2    G-----SD..T.........SA.....Q...........QP.S..........SKADN..
psE2    ENT...GFA.DP.APG....SKKEDPS..LT..Q.Y.L.DSSS-.....IS.IDTTKVAS
psA2    FEQS--SN..S.........SAKTAE...LT..E...LA.T.E-...K.IS.E.GNDS.Y
psA4    ----------------------.----------------..----------.--
psA3    FNT-.DKTCTNK.D.T.K..SA.TDANR.LT.M..Y.PIDSTDQQGKK..P..SVDDK.R
psC3    FNT-.DKTCTNK.D.T.K..SA.TDANR.LT.M..Y.PIDSTDQQGKK..P..SVDDK.R
psD2    ..AS.N.A..T...E.....AKDTK..Q..T..Q.Y.L.DSSS-.....IS.SDTANG.R
psD3    ----------------------.----------------..----------.--
psC4    ----------------------.----------------..----------.--
psB1    D..SFVCEL---------------.----------------..----------.--
psD1    LET.---TCTKK.ADT..K..AETDIN...E.M..Y.L.TE.V...K...P.SDTAKG..
        LET.---TCTKK.ADT..K..AETDIN...E.M..Y.L.TE.V...K...P.SDTAKG..

CRP72   D---GCAECTKEST-GPLKCTKCKPNRKPAG--TSDNYTCTEKTCENPTACGGTAGSCDA
pQ1     E.....SA.SSN--..AF...D....Y.KE..........-V....DE......S.A...
psB2    .......T..FS--..S.T.NS....Y.-QN..----------..----------.S.A...
psA1    E.........GTT-..S.............N.-Q...........DE............
psC2    E.....SA..ND--G.AF...D....YRKE...-.TGSVTCT....DD.T....S.A...
psE2    R.........NS--..TF......V.YR.S.E.-PSTGVTCT.V..D.......S.A...
psA2    SGLAR.LS..APAAP..AS.NR..IGY.LQ-..-----G.TCV....DE......S.A...
psA4    -...----------.----------..----------..----------
psA3    E.....SV.SNN--..GF...E..V.Y.KQSNGDAGDDYTCV....D.........A...
psC3    E.....SV.SNN--..GF...E..V.Y.KQSNGDAGDDYTCV....D.......S.A...
psD2    ......SA.SNS--..GF...D....Y.KQLNGDAGDDYTCT....DE.......A...
psD3    -...----------.----------..----------..----------
psC4    -...----------.----------..----------..----------
psB1    ......SV.SND--GNTS...D....YRKQSD.S.PGSVTCT....D.......S.A...
psD1    ......SV.SND--GNTS...D....YRKQSD.S.PGSVTCT....D.......S.A...
```

```
CRP72   IVIDDQGTTKHYCSYCGDSSQAPIDGLCASEAQKAG-NT-CANGVCTQCT-NNYFLYMGG
pQ1     ....EN.N......F..E.GKF........D--..NN.G........S..AA........
psB2    MI..............EANKF....I.T.DNN.K.T....DSHT....A.QG........
psA1    T...N..NA........EANKF..N.V.VESS..N...N..NSHT..S.A.QG..M....
psC2    MI.............................TD--.N......T.N...S...QG.....D.
psE2    ....NT.KEL.......KD.EF....X.....--K....G.V.N...S...MG.......
psA2    .....K.N............KY...........---S....N.H..ES...TG.....N.
psA4    ------------------------------------.--.----------.---------
psA3    MI.....N.........ETNKI....K.VDS-GSING....NSHT..S.A.A........
psC3    ....ND.VEH.......NQGDV..N.I.T.S---LAS.....D...KS.A.QG..M....
psD2    ..VGND.SMLS...K.VGAGYG..N.K.TN---ALAG.....D....R............
psD3    ------------------------------------.--.----------.---------
psC4    ------------------------------------.--.----------.---------
psB1    MI...K.NA........ETNKF....I.N.D..---S....N.H...S...TG.......
psD1    MI...K.NA........ETNKF....I.N.D..---S....N.H...S...TG.......

CRP72   CYSTQKAPGSFMCKTAGNTGICTE-AANNRYFVVPGASNTDQSVLACSNPLGTLTGTGDT
pQ1     ..KVNTV...H.....-.N.V..AVSE..K..I......QN......G.....EL----.
psB2    ..KVGQE.......ASSSN........S.K..L.............................
psA1    ..DAS....NH.....D.-.V..TPN......L.....................V......
psC2    .....SQ..NL.....N.-.V..AVNE..K..I..E.KP.Q.......G......VD.---
psE2    ...ISAQ..KS..TK..D-.V....-.AAG..IP.SPTKDK....S.G....VELAGQ--
psA2    ..DVS.P..NL..SK.TTA.V........K.....E.KA........G......I..QG.
psA4    ------------------------.-------------------------------------
psA3    ..KATEV...L.....D.-.V..AAN...K.......T.QN......G......I..QG.
psC3    ..DVS.....H...K.D.N........S.........QT........G......V..QG.
psD2    ..KATEV...L..SE.TTA.V..TPN..D......E.KA........G......VDPQG.
psD3    ------------------------.-------------------------------------
psC4    ------------------------.-------------------------------------
psB1    ...VSAQ..NY.....DSS.............L.............................
psD1    ...VSAQ..NY.....DSS.............L.............................

CRP72   AKAYVGVEGCSQCTAPAALSDGGMAPAVCTSCDS-------------------------
pQ1     .......K-------------------------.............................
psB2    .......D........T..TE....A.....................................
psA1    .......LN..............A.I.....................................
psC2    -......D........T.P.E...T.....................................
psE2    -......D........T.P..A..T.....................................
psA2    .......D........TQ.EAP..TS....A...............................
psA4    ------------------------------.-------------------------------
psA3    ..HTL.MD........T..TTA..TA.I..A...............................
psC3    .......D.........G..E...T........N............................
psD2    .......D........T..TE....A.....................................
psD3    ------------------------------.-------------------------------
psC4    ------------------------------.-------------------------------
psB1    .......Y...A....TQ.EAP.....T..A.IGNNKPNLAGSGCFACTVSGCSHCGAGD
psD1    .......D...A....TQ.EAP.....T..A.SGNNKPNLAGSGCFACTVSGCSHCGAGD

CRP72   --------SKKPNRDGSGCVLCSVGGCKSCVMDNICGECNSGFSLDNGKCVSSGANRSGL
pQ1     ........-------------------------------------------------------
psB2    ........DR..................N.............................T.....
psA1    ..........V.............DTN.......GV....SD.Y..EG.R.............
psC2    ........G.......T...T..DTN....A..GV..................T.........
psE2    ........DR...K......T....D.......DV....SD.Y..EG................
psA2    ........G..............DTN.......SV..............................
psA4    ........-------------------------------------------------------
psA3    ........G................D.................................S.....
psC3    ............K.........DTN.......GV..............................
psD2    ........DR.....................V.GV..............................
psD3    ........-------------------------------------------------------
psC4    ........-------------------------------------------------------
psB1    KCEGCTSKDQR.SL...Q.IA..ID..VR.SEE.K..Q.SD.YR.EG.R.............
psD1    KCEGCTSKDQ..SL...Q.IA..ID..VR.SEE.K..Q.SD.YR.EG.R.............
```

```
CRP72   SAGAIAGISVAVVAVVGGLVAFLCWWFVCRGKA
pQ1     -----------------------------------
psB2    .T..........V......G.F....I.....
psA1    .T..........I......IG............
psC2    .T..........V......G.............
psE2    .T..........V......G......L.....
psA2    .T..........I.....F.G............
psA4    -----------------------------------
psA3    .T..........I......IG............
psC3    .T..........I.......G............
psD2    .T..........I.......G............
psD3    -----------------------------------
psC4    -----------------------------------
psB1    ..................G.............
psD1    ..................G.............
```

This appeared clearly not to be the case. The comparative examinations of the nucleotide and inferred amino acid sequences supported the conclusion that the upstream segments missing from these different pseudo genes have been lost as a result of recombinational deletion(s).

In comparison to many variable residues that can be identified in the aligned polypeptides (**Fig. 4.27**), some amino acid positions show definite evidence of conservation. Three features deserve comment:

a) At positions where substitutions have occurred, some residues are common to several of the proteins. Other replacements are restricted to two or three proteins. For example, the amino acid sequences specified by psB1 (pseudo gene B1) and psD1 (pseudo gene D1) are identical except for one residue. The sequences specified by psA3 and psC3 also exhibit a substantial number of identities. It appears from the alignment of these inferred amino acid sequences that mosaic recombinations of small segments might have occurred between various *vsp*72 subfamily loci and resulted in the formation of these loci.

b) Cysteine tetramer (CXXC) motifs are conserved between all 13 pseudo VSP, CRP72 and pQ1 (where overlapping sequence data are available), both in number and location. An exception is the psB1/psD1 pair, for which three additional CXXC motifs are evident within a portion of the C-terminal segment. This latter appears to be encoded by an 'insert' that is unique to the pcB1 and psD1 ORFs.

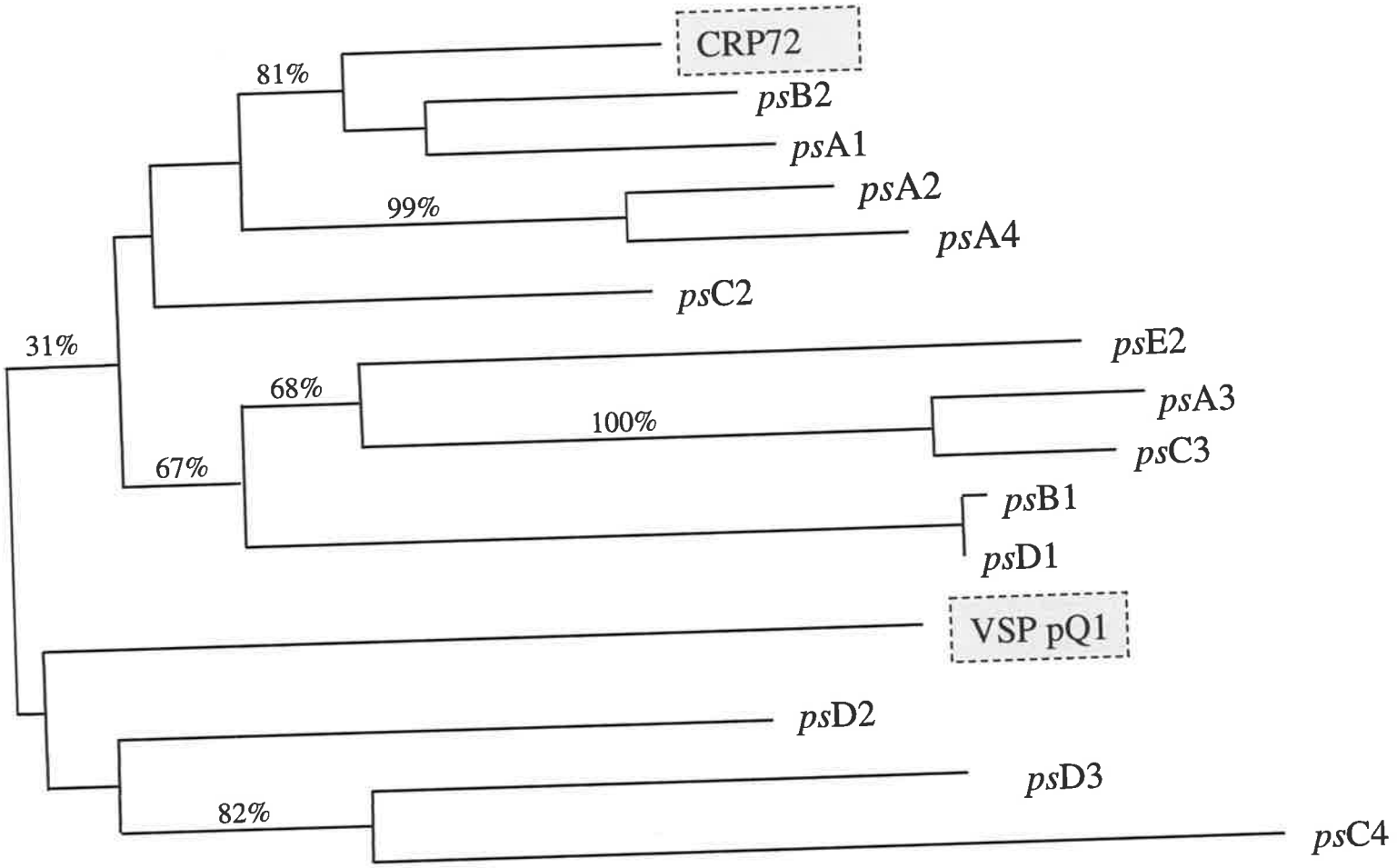
c) Like all previously inferred VSP, the polypeptides inferred by the 13 *crp*72-like pseudo genes contain the highly conserved C-terminal hydrophobic (transmembrane) and invariant (-CRGKA) domains. The parent (pseudo)genes also possess apparently intact 3' termini, including appropriate stop codons and downstream polyadenylation signal sequences.

The aligned amino acid sequences were subjected to Neighbour-Joining analysis using MEGA. This yielded the phylogenetic tree shown in **Fig. 4.28**. Despite minor problems caused by comparison of sequences that differed in length and the region of overlap, analyses using sequence subsets (the 5', central, or 3' segments) or different parameter settings (complete- or pairwise-deletion) give very similar results. Consequently, the broad relationships revealed in **Fig. 4.28** were well supported. The sequences show differing degrees of similarity, with several distinct subsets apparent. These reflect the most closely related sequences evident in **Fig. 4.27** and based on the near-identity of some pairs, e.g. psA2 & psA4, psA3 & psC3, psB1 & psD1, it seems likely that they represent recently duplicated loci. It is difficult to assess how frequently the various loci that comprise this *vsp72* subfamily undergo duplications and/or recombination or how long the pseudo genes have been (or remain) nonfunctional. However, the identification and similar structure of multiple 'paired' loci (such as those just mentioned above) suggests that these loci were duplicated as pseudo genes. Whether the duplications represent *gene* duplications or partial chromosome replications remains unknown.

Some of the deduced pseudogenes are obviously very closely related to CRP72 (**Fig. 4.28**). Other sequences, although still clearly related to CRP72, are more divergent as evidenced by their deeper branching patterns. The pQ1 sequence (VSP21pr199) is among the more divergent members of the subfamily, but inspection of the aligned sequences (**Fig. 4.27**) shows that there are, nevertheless, a considerable number of amino acid identities between this polypeptide and CRP72 across the entire length of the alignment. Thus, even those polypeptides that form the deepest branches of Fig. 4.28 may be considered unambiguously to be members of the 'VSP72' subfamily.

The aforementioned results, based on hybridisation and sequence analysis of 18 unique cloned genomic DNA *Sac* I restriction fragments that hybridised strongly with probes

Figure 4-28. Inferred relationships between CRP72 and the various VSP72-like polypeptides specified by the functional (VSP pQ1) or pseudo (*ps*) genes described in this chapter. Different segments of the aligned amino acid sequences shown in Fig. 4-27 (483-, 648- or 753-residue overlaps) were subjected to phylogenetic analysis by the Neighbour-Joining method using MEGA (Kumar *et al.* 1994). Analyses were undertaken using the γ-2 distance measure. Results obtained using pairwise or complete deletion options were very similar, the only major differences being the alternate placement of *ps*C2 with either the *ps*A2/A4 or *ps*A3/C3 clusters and of *ps*D2 with either the CRP72 or pQ1/*ps*D3/C4 clusters. Confidence values, determined by bootstrap analysis from 5,000 iterations, are indicated.


CRP72
81%
psB2
psA1
psA2
99%
psA4
psC2
31%
psE2
68%
psA3
100%
67%
psC3
psB1
psD1
VSP pQ1
psD2
psD3
82%
psC4

(M165-[5'], M165-[3']) derived from a *crp*72-like gene segment, indicated that multiple vsp genes (including a large number of pseudo genes) closely related to the *crp*72 gene are present in the Ad-1/c3 genome. Sequence analysis of five inserts resulted in the identification of 13 vsp pseudo genes arranged in head-to tail tandem arrays, and partial sequence analysis of a sixth insert revealed a 1.6-kb segment of a putative functional vsp gene (also closely related to *crp*72). Limited analysis of other *Sac* I genomic inserts by PCR indicated that most contained several *crp*72-like genes. On the basis of these findings the Ad-1/c3 genome appears to contain a large subset of genes (the '*vsp*72' subfamily) that are closely related to *crp*72. Unfortunately, the data provide an incomplete picture of this gene subset. In particular, they give no clear insight into the absolute number (or ratio) of pseudo and functional genes within the subfamily because the demonstrated prevalence of pseudo genes in only 5 of >20 related genomic restriction fragments may present a highly biased view of the complete subfamily.

4.12 Detection of variants expressing *vsp*72-like genes in a culture of *G. intestinalis* trophozoites

It was of interest to know whether the *vsp*72 subfamily is comprised mainly of pseudo or functional genes and how frequently variants express the latter within heterogeneous cultures. This was examined by *in situ* mRNA hybridisation. If a large number of functional *vsp*72 genes exist within the genome, the frequency of cells staining with *vsp*72-specific anti-sense probes may be higher than that of cells hybridising with probes that are known to be specific for transcripts from a defined single vsp gene locus. Furthermore, by making several probes, some specific for functional '*vsp*72' subfamily genes and others specific for both functional and pseudo '*vsp*72' subfamily gene sequences, there seemed to be potential for investigating whether some of these pseudo genes are expressed, perhaps through recombination or gene conversion, into functional, expressed loci.

Combinations of PCR primers were used to amplify from Ad-1/c3 genomic DNA (containing the full spectrum of vsp genes) or *Sac* I plasmid constructs (containing known '*vsp*72' subfamily inserts) segments that corresponded to different regions of known '*vsp*72' subfamily genes. The amplified fragments were gel-purified and used as templates to prepare DIG-labelled, single-stranded probes using the appropriate reverse (anti-sense) or forward (sense) primers. Some of the probes for the more heterogeneous known segments of '*vsp*72' subfamily genes were mixed, in order to identify all the cells expressing loci with related sequences. The primer combinations used to prepare templates for probes that were putatively specific for functional genes were:

1. Oligos 199 + 200
2. Oligos 193 + 201, or 193 + 202
3. Oligos 194 + 201, or 194 + 202

Antisense probes made from these afore-mentioned products were used to identify cells expressing functional '*vsp*72' subfamily genes. Other primer combinations were used to prepare templates for probes that were putatively specific for segments identified mainly in the *vsp*72-like pseudo genes (but also in some functional genes):

4. Oligos 195 + 204, or 196 + 204
5. Oligos 195 + 178
6. Oligos 196 + 203

For determining the frequency of cells expressing a single-locus vsp gene, anti-sense and sense (negative control) probes specific for *vsp*417-6 were used. The *in situ* mRNA hybridisation procedure is described in detail in section 2.14. Anti-sense and sense (control) probes were each tested for hybridisation on trophozoites from a long-term culture of the Ad-1 isolate. After staining, the numbers of cells hybridising with each anti-sense probe were counted (minimum total cell count: 10,000) and the frequency of cell variants containing

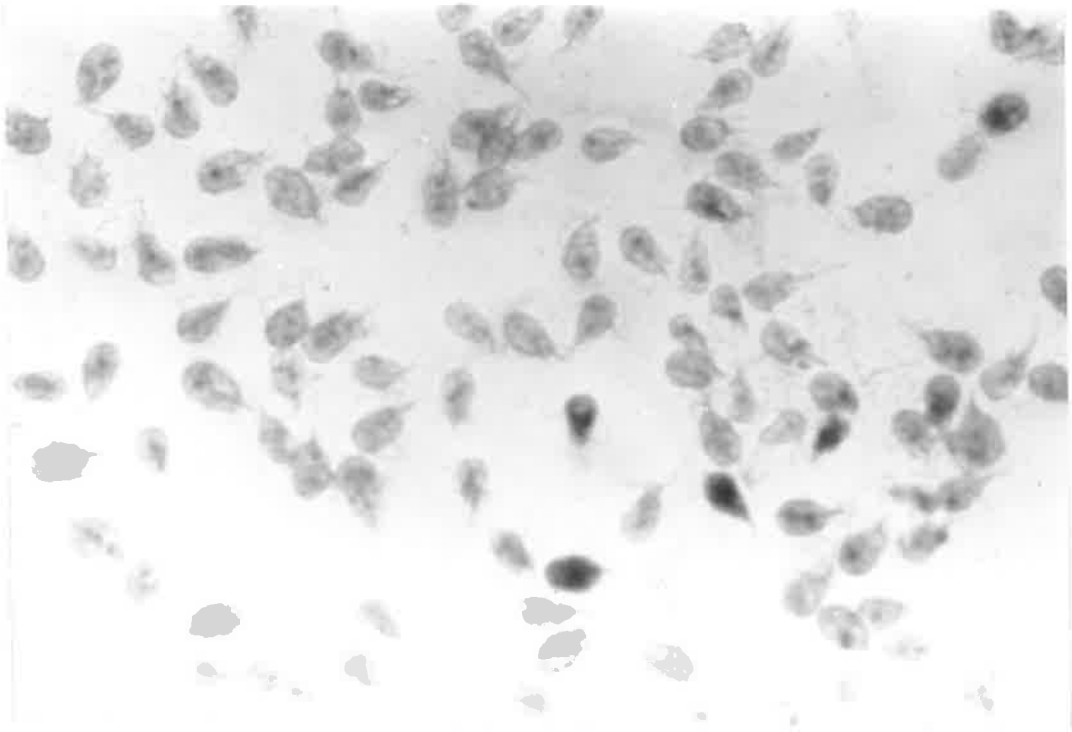
transcripts containing a sequence detected by each particular probe was calculated as a percentage of the entire trophozoite population. The data are summarised in **Table 4. 10**.

As is apparent from **Fig. 4.29**, more cells (6.45%) were stained by some anti-sense probes, e.g. prepared from the oligo (199 - 200) template, than by other anti-sense probes, e.g. prepared from the oligo (195 - 178) template (< 0.06% stained). The 'sense' probes were uniformly negative, i.e. none produced any cytoplasmic staining as exemplified in **Fig. 4. 29**. The most frequent variants were those detected with the anti-sense probes corresponding to the oligo (194 - 201/202) templates. These comprised 8.39% of the trophozoite population. This was not surprising, as the oligo (194 - 201/202) segment seems to be relatively conserved between the different '*vsp*72' subfamily loci that were described in the earlier sections (**Fig. 4.29**). However, using the anti-sense probe prepared from the oligo (199 - 200) template, which should detect *crp*72 and other, closely related genes belonging to the '*vsp*72' subfamily, 6.45% of trophozoites were stained. This indicated that a relatively large number of '*vsp*72' subfamily loci may be available for expression by variant trophozoites within cultures of type A-I *Giardia*. Alternatively, the modified TYI-S-33 medium or other conditions of *in vitro* culture may have favoured the growth of cells that express these particular *vsp*72 genes. Recently published evidence supports this latter possibility (Singer *et al.* 2000).

Figure 4.29. Detection of *vsp72* subfamily gene expression in variant Ad-1/c3 *Giardia intestinalis* trophozoites by *in situ* mRNA hybridisation. DIG-labelled, single-stranded 'antisense' probes, specific for sequences within the 5' portion of functional and/or pseudo *crp72*-like coding sequences (Fig. 4.24, Table 4.10), were used to identify cells that contained related transcripts (see M&M, section 2.14). Specificity controls included samples exposed to the complementary 'sense' probes. The examples shown include:

Plate (a). ***Negative control, using a 'sense' probe with specificity for both functional and pseudo genes.*** The single-stranded probe was synthesised using primer 194 from a double-stranded template that had been produced in PCR using primers 194 and 201/202 (c.f. Fig. 4.24). It should not have hybridised with mRNA. No stained cells were detected.

b



c

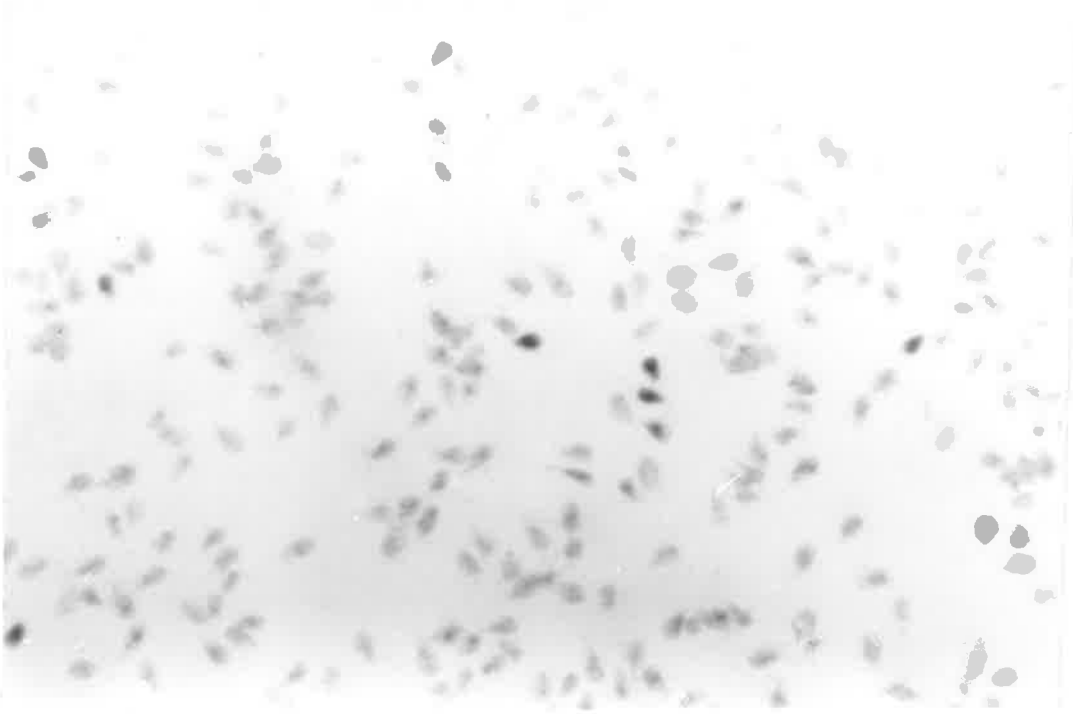


Figure 4.29. Detection of *vsp72* subfamily gene expression in variant Ad-1/c3 *Giardia intestinalis* trophozoites by *in situ* mRNA hybridisation. DIG-labelled, single-stranded 'antisense' probes, specific for sequences within the 5' portion of functional and/or pseudo *crp72*-like coding sequences (Fig. 4.24, Table 4.10), were used to identify cells that contained related transcripts (see M&M, section 2.14). Specificity controls included samples exposed to the complementary 'sense' probes. The examples shown include:

Plate (d). ***'Anti-sense' probe specific for functional genes only.*** This probe was synthesised using primer 200 from a double-stranded template that had been produced in PCR using primers 199 and 200. Based on all available nt sequences, this segment is present in functional genes but not pseudogenes (c.f. Fig. 4.24). Heavy cytoplasmic staining was observed in a minority of cells. Of this cell population, 6.45% stained as a result of hybridisation with this anti-sense probe (Table 4.10).

d

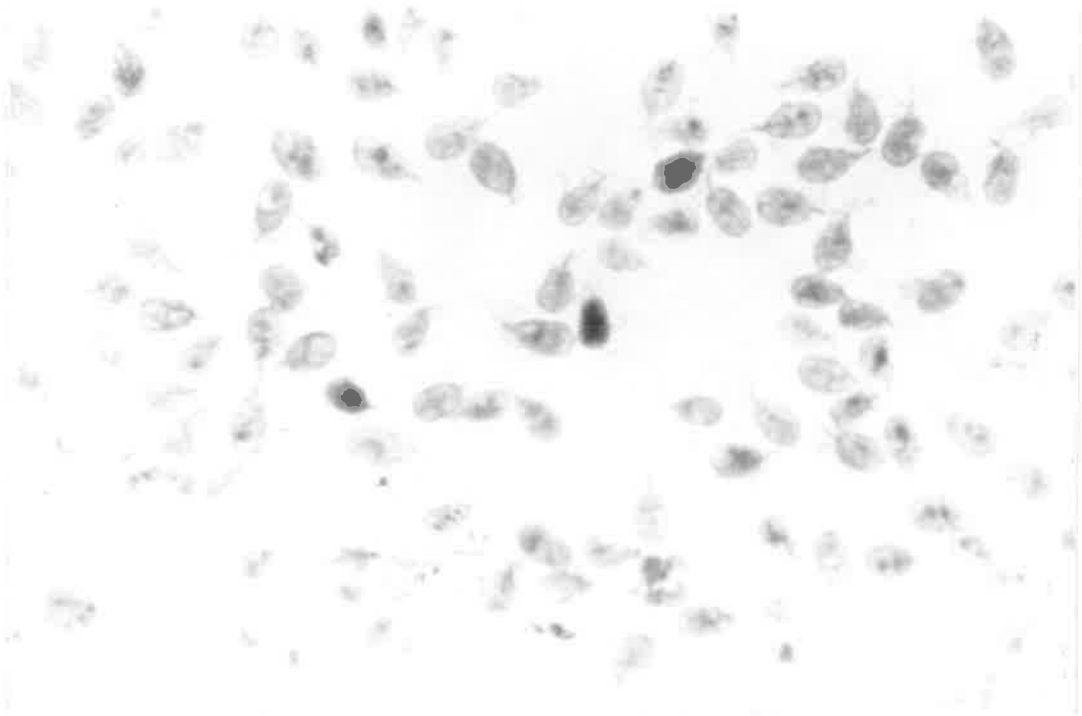


Table 4.10. Incidence of Ad-1 trophozoites stained by *in situ* mRNA hybridisation using *vsp* 72 anti-sense probes

DNA template[a]	Stained cells/total[b]	Percentage[c]
199 + 200	786 / 12,180	6.45%
194 + 201/202	856 / 10,200	8.39%
195/196 + 204	580 / 10,800	5.37%
195 + 178	6 / 10,000	0.06%
196 + 203	99 / 12,000	0.82%
vsp417-6	150 / 18,900	0.80%

[a] Oligos 178, 200, 201, 202, 203, 204 were used to make antisense probes from corresponding DNA templates fragments.

[b] The number of stained cells observed relative to the total number of cells counted.

[c] The percentage of trophozoites stained for the respective probes.

4.13 Discussion

The complete sequences of about 20 apparently functional giardial vsp genes are now available in the published literature and GenBank. In contrast, there are almost no data on vsp pseudo genes, except for the single 693-bp sequence described by Upcroft *et al.* (1993b) which was corrupted by multiple stop codons throughout the pseudo gene reading frame. Other sequences, evident in the *Giardia* Genome database, may represent similarly corrupted vsp genes but these single-run sequences have yet to be corrected, compiled and subjected to a focused analysis.

The major findings of this chapter are the definition, within the *Giardia* genome, of a large vsp gene subfamily related to the *crp72* gene described by Adam *et al.* (1992b). Most of the characterised loci exist in a non-functional form which consists of an uncorrupted open reading frame lacking only a segment at the 5' end that would normally encode the first (N-terminal) 70-105 amino acid residues of putative ≈700-residue polypeptides. These were all organised as tandem gene arrays in close, head-to-tail arrangements. The identification of such pseudo genes that are essentially intact, except for the missing 5 ends, and their arrangement within such tandem gene arrays, are novel findings. At present, only two published examples exist of loci existing in the neighbourhood of a vsp gene. The first example is the tail-to-tail arrangement of two identical copies of the *vsp1267* gene, which are separated by an intergenic region of approximately 3 kb (Mowatt *et al.* 1991). The second example is the description by Upcroft *et al.* (1997) of two gene arrays located near the telomeric rDNA repeats. Both arrays consist of a single vsp gene (*crp136* or *crp65*) lying in a head-to-head arrangement with two non-vsp gene loci, encoding a protein kinase (PK1 or PK2) and ankyrin homologues (ANK1 or ANK2). The distance separating each vsp gene from its PK neighbour was approximately 700-900 bp. These examples are clearly very different from the tandem pseudo gene arrays discovered in this PhD project.

In addition to pseudo genes, 1.6 kb of sequence was obtained which appears to represent a functional locus belonging to the *vsp72* gene subfamily. The characterisation of this locus is incomplete, as the work was done during the final months of the project. However, preliminary screening (by PCR) of the various cloned genomic fragments identified several other loci that seem likely to represent functional *crp72*-like genes. Additional putative pseudo genes that may be organised in extra tandem arrays were also identified during this final screening (**Table 4.9**). Most of the characterised *vsp72* subfamily loci are pseudo genes. However, these are likely to represent only a small fraction of the whole subfamily and the ratio of functional to pseudo genes has yet to be determined.

The cloning of at least 18 unique genomic DNA restriction fragments, all containing one or more coding sequences with similarity to the *crp72* gene reported by Adam *et al.* (1992b) and some containing additional loci encoding other VSP, was an important step in distinguishing unambiguously a number of very similar *crp72*-like pseudo genes (**Table 4.2**). The distinction between inserts was complicated by their content of multiple, highly related vsp gene sequences. This, together with the very large size (mean 7.6 kb; maximum 13 kb) of many of the inserts necessitated an arduous mapping program to eliminate duplicate clones and to distinguish unique fragments of similar size. Although restriction data were obtained for all of the cloned fragments, subsequent work was focussed on five inserts chosen randomly from the panel. These were subjected to a detailed characterisation, involving further restriction and Southern hybridisation analysis and nucleotide sequence determinations involving numerous subclones (section 4.8). The subcloning of various fragments from the parent constructs was necessary because the presence of the tandem gene arrays in the respective inserts made it impossible to determine sequences internally by primer-based reactions or PCR-generated segments using the original clones.

The vsp gene loci described in this chapter all belong to the *vsp72* gene subfamily. Their similarity is evident from the deduced amino acid sequence alignment shown in

Fig. 4.27 and also from the results of phylogenetic analysis shown in **Fig. 4.28**. However, as can be seen from both of these figures, some loci (e.g. *ps*C4 and *ps*E2) appear to differ substantially whilst others (e.g. *ps*A2 and *ps*A4) are more closely related. The sequence differences (and similarities) evident in **Fig. 4.27** show many conserved positions throughout the length of the inferred polypeptides. However, it is unclear whether these are the result of multiple intergenic recombinations or whether they reflect accrued point mutations that occurred during the period in which these loci were functional. In the latter case, the mutations would have been subjected to natural selection. The surprising integrity of these pseudogenes, i.e. the absence of internal stop codons which could be expected to arise over time, suggests either that these loci have become nonfunctional and replicated only very recently, or that they are still functional (and thus subject to selective pressures) by contemporary recombinations and gene conversion as discussed in the next chapter.

Chapter 5

Discussion

In this study, 18 loci that define a new subset of vsp gene sequences (designated the *vsp*72 subfamily, based on the prototype gene *crp72* (*vsp1269*) described previously by Adam *et al.* 1992) have been identified. A major novel finding was the discovery of numerous *crp72*-like pseudo genes, organised in tandem gene arrays. A related, putative functional *crp72*-like gene was also partially characterised. A second vsp gene subset (designated the *vsp136* subfamily) was also defined by the discovery and characterisation of five genes containing tandem repeat elements. These were similar to the *crp136* gene described by Chen *et al.* (1995), which together with the *crp65* gene (Chen *et al.* 1996) had been hypothesised to belong to a gene subset (Upcroft *et al.*, 1997). Three genes contained tandem repeats that were identical to those found in *vsp136*-1 (*crp136*), one (*vsp136*-2) possessing, like *vsp136*-1, 23.5 copies of the repeat unit and the other two, *vsp136*-3 and *vsp136*-4, possessing 1.5 and 20.5 copies of the repeat respectively. In total, more than 43,568 bp of nucleotide sequence data encompassing these and other vsp genes and their flanking regions, derived from cloned genomic fragments and PCR-amplified DNA, were submitted to GenBank.

The importance of these findings is not restricted to the identification and characterisation of 23 additional vsp gene loci in the *G. intestinalis* genome. Except for the *vsp*417 gene subfamily which has been studied in detail in genetically distinct isolates (Ey *et al.* 1998, 1999 & unpubl. data), little is known about the size, stability or nature of vsp gene subsets. Indeed, few of the polypeptides encoded by completely characterised vsp genes show extensive similarity with one another. It is only by comparing loci that belong to the same or closely related subsets that a deep understanding of their evolutionary history and structural similarities can be obtained, together with data pertaining to the mechanisms involved in VSP switching.

The project began with an initial aim of studying the expression of *vsp52* in the clonal *G. intestinalis* line, Ad-1/c3, because this gene had been identified in a genomic DNA library using an antiserum raised against the surface protein complex from these cells. However, RT-PCR and *in situ* mRNA hybridisation experiments carried out in the early stages of the project showed that the presumption that VSP52 was produced by the majority of Ad-1/c3 trophozoites was incorrect. The project was therefore directed into the study of vsp gene subsets, which had been one of the original aims. This area of work proved highly productive.

The *vsp*136 subfamily, one of the two vsp gene subfamilies studied in this project, can now be defined by eight apparently functional loci. The five newly described genes contain tandem repeat sequences that are identical (*vsp136*-2, -3 and -4) or similar (*vspR2*) to the *vsp136*-1 (*crp136*), prototype of this subfamily but with different tandem repeat copy numbers as mentioned above. The two other known members of this gene subfamily, *crp65* (Chen *et al.* 1996) and *vsp52* (Ey *et al.*, unpubl. data), contain tandem repeats that are very different from the *vsp136*-1 repeat. However, all of these genes encode very similar non-repeat N- and C-terminal amino acid sequences (**Fig. 3.16**). These differences in repeat copy number may have arisen by intragenic deletions (insertions), or by intergenic recombination between these or other as-yet unidentified genes that contain the same repeat element. Examples of putative alleles that contain different numbers of the same repeat element have been described previously for two vsp genes, *vsp*A6 (*crp170*; Yang & Adam 1994; Mowatt *et al.* 1994) and *vsp*C5 (Yang *et al.* 1994). In the case of the *vsp*A6 gene, three alleles were identified: the expressed allele, *vsp*A6.1, containing 18-23 copies of the repeat, and two non-expressed alleles, *vsp*A6.2 and *vsp*A6.3, containing 9 and 8 repeats respectively (Yang & Adam 1994). Mowatt *et al.* (1994) studied cloned lines from different isolates of *G. intestinalis* for variants producing VSP that reacted with mAb 6E7 (which is specific for an epitope encoded by the 195-bp *vsp*A6 repeat). They observed differences in the size of the

transcripts from different 6E7 mAb-reactive clones that were explained by differences in repeat copy number. In the case of *vsp*C5, multiple alleles containing different numbers (ranging from 10 to 26 tandem copies) of the 105-bp repeat were identified within the genomes of two subclones of the WB isolate (Yang *et al.* 1994).

The extension of nucleotide sequence identity across the 5' and 3' flanks of the *vsp136* loci may have important implications. These nearly identical flanking sequences may facilitate gene rearrangements that translocate the vsp gene loci into possible expression sites (like VSG expression in African trypanosomes; Bangs *et al.* 1997) or enhance recombinations that promote DNA breakage and repair. In fact characterisation of vsp genes which contain identical repeat sequences with different copy numbers of repeats is another indication of recombination events which resulted in gaining or losing the repeat sequences in these genes. It is possible that the 5' flanking sequences of these genes contain promoters, but such regulatory elements have not yet been identified in giardial vsp genes.

From an examination of the aligned amino acid sequences shown in **Fig. 3.16**, it is possible to gain an insight into the way in which some of the identified *vsp136* subfamily genes may have evolved (**Fig. 5.1**). As illustrated earlier in **Fig. 3.17**, the C-terminal segments show evidence of three major gene lineages:

(a) *Vsp136*-1 and *vsp136*–2. These are essentially identical across their entire coding sequences, but differ only by apparent breakpoints in the untranslated flanks (**Figs. 3.18 & 3.19**).

(b) *Vsp136*-3, *vsp136*-4, *crp65, vspR1* and *vsp52*. These are nearly identical across the C-terminal segment (**Fig. 3.16**), with the exception of *vsp52* which is more divergent. Within this segment, this whole group differs by multiple substitutions from group (a). The repeat units of both *crp65* and *vsp52* differ from the identical

repeat units of *vsp136*-3 and *vsp136*-4 (these are the same as the *vsp136*-1 repeat). The repeat unit of *vspR1* has not been characterised.

(c) *VspR2*. Within the C-terminal segment (**Fig. 3.16**), VSPR2 shows more similarity to group (b) than to group (a). A number of common residues are shared with the group (b) sequences. The repeat unit is very different from those of groups (a) and group (b).

As is depicted in **Fig. 5.1**, the evidence is consistent with an early duplication of a single ancestral gene (A) to yield two genes, B1, the precursor of group (a), and B2, the precursor of groups (b) and (c). This duplication must have occurred sufficiently long ago to have allowed the accumulation of the many C-terminal sequence differences that distinguish the group (a) sequences from those of groups (b) and (c), as evident in Fig. 3.16. Subsequently, a second duplication of the B2 locus must have occurred, to yield B2*a* (with subsequent duplications giving rise to B2*b*, B2*c*, etc.) leading to group (b), and C leading to *vspR2*, i.e. group (c). This also is predicted to have occurred sufficiently long ago to have allowed the accumulation of the mutations that distinguish group (b) sequences from *vspR2*. The gene duplications that yielded the various members of group (b) are predicted to have occurred very recently, as the encoded C-terminal sequences of these VSP are essentially identical, i.e. there has not been time for mutations to emerge. With respect to *vsp136*-1 and *vsp136*-2, their near identity suggest two obvious possibilities. The first is that they represent the same locus, with the recombinations in the flanking segments occurring in one of the progeny (WB or Ad-1) from which these sequences have been obtained. The second possibility is that they do represent distinct loci, which have arisen via a very recent gene duplication. In this latter case, the recombinations in the flanking segments would have occurred in one locus but not the other. The overall paucity of mutational differences between these genes, even at synonymous ('silent') nucleotide sites, supports a relatively recent replication.

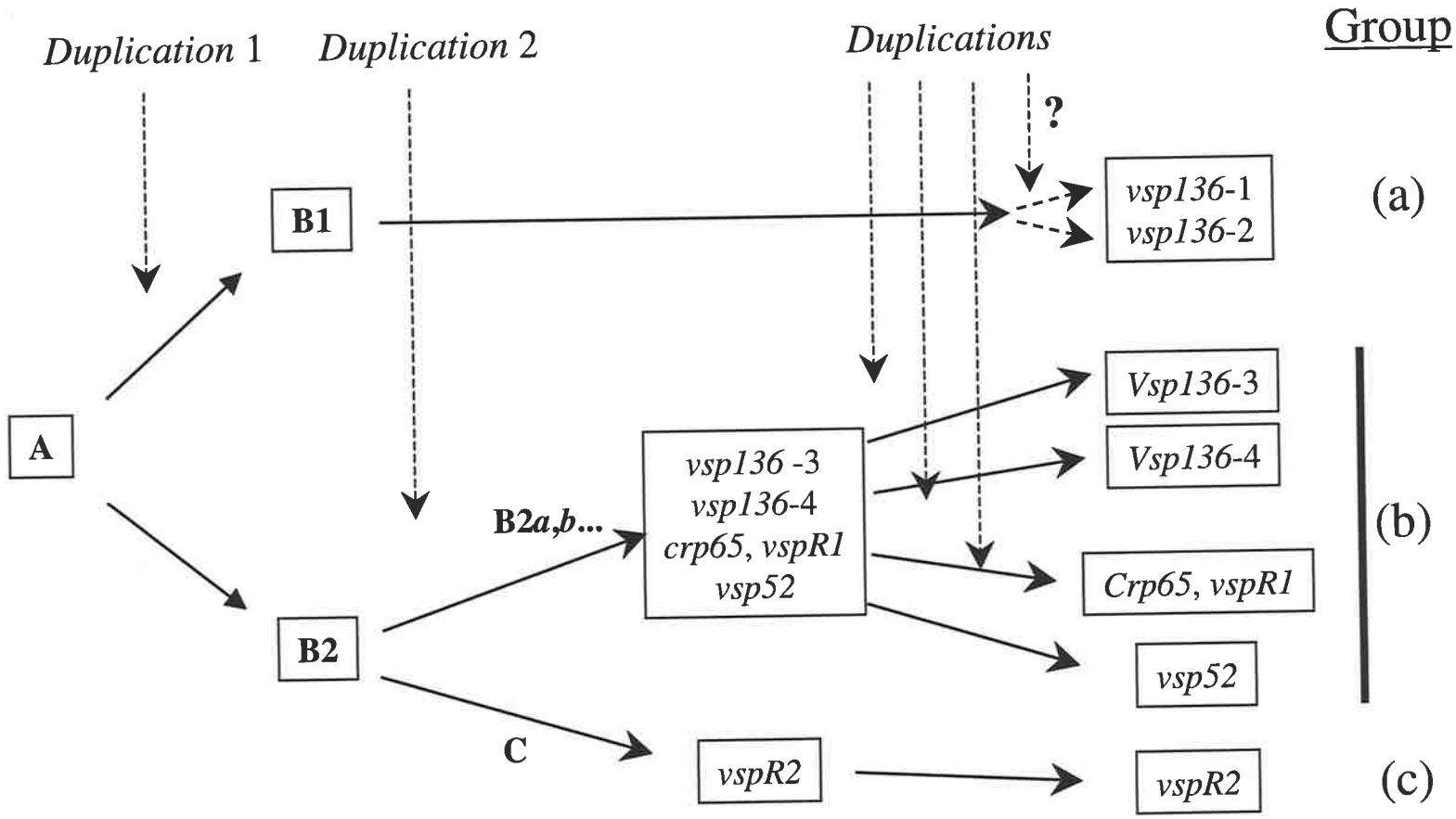


Figure 5.1. Hypothetical evolution of the *vsp136* gene subfamily. Loci are defined on the basis of the encoded C-terminal non-repeat sequences. A duplication of ancestral gene A is predicted to have yielded identical loci B1 and B2. Locus B1 is the contemporary *vsp136*-1 *(crp136)* gene, which may have duplicated very recently to yield *vsp136*-1 and *vsp136*-2 (group a). Locus B2 subsequently duplicated to give rise to two distinct gene lineages, one leading to the various group (b) loci and the other leading to the contemporary *vspR2* locus.

As discussed above, the detection of these different *vsp*136 subfamily genes that possess identical or very similar flanking and coding sequences indicates that they are the result of serial replications of an ancestral gene. Some of these copies show evidence of recombinations, e.g. the breakpoint of *vsp136*-2 approximately 240 bp upstream from its start codon (**Fig. 3.18**) and in the 3' flanking segments of other loci (**Fig. 3.19**). It is not known whether the observed identity between the 5' flanking and coding regions of the *vsp136* subfamily genes represents conservation (due to natural selection) or an insufficient period of time (since the replication events that gave rise to the multiple loci) for many mutations to accumulate

The findings described in Chapter 4 for the *vsp*72 subfamily, which lack tandem repeats, stand in stark contrast to those discussed above for the *vsp136* subfamily. Several interesting features about the characterised members of this gene subfamily deserve mention:

- A surprisingly large number of related sequences were detected within the genome by Southern hybridisation analyses. These can be considered, on the basis of the moderate stringency used for the hybridisations, to belong to this subfamily. Many of these sequences were cloned within genomic (*Sac* I) restriction fragments and subjected to partial or complete (i.e. nucleotide sequence) analysis.

- A large number of pseudo genes was detected. This was the first time that vsp pseudo genes have been described in *Giardia* with intact entire open reading frames except the 5' initiation codon and the immediate nucleotide sequences (which encoding the signal peptide segments of the VSP).

- All of the pseudo genes identified in these genomic fragments were arranged in tandem arrays, in head-to-tail arrangements

- A segment of what appears likely to be second functional member of this gene subfamily was identified, the prototype *crp72* [*vsp1269*; Adam *et al. 1992*] gene being the only other known functional locus.

Alignment of the hypothetical polypeptides specified by the various *vsp72* pseudo gene sequences (**Fig. 4.27**) revealed many more differences than were evident between the VSP encoded by the characterised members of the *vsp136* subfamily discussed above (**Fig. 3.16**). Nevertheless, it was apparent from both the amino acid sequence alignment (**Fig. 4.27**) and the phylogenetic tree (**Fig. 4.28**) that certain sequences within the characterised panel were substantially more related to each other than to the other sequences. This is evident in **Fig. 4.27** by the similar sequence motifs of psA2 and psA4, psA3 and psC3, and psB1 and psD1, which were identified as closely related pairs in **Fig. 4.28**. Some of these genes have been characterised only partially by nucleotide sequence analysis because they were truncated by the restriction cleavage used in cloning the fragments.

Preliminary data on the remaining cloned fragments showed that there are many more loci present within these inserts, as well as additional uncloned genomic fragments, which remain to be identified. The characterisation of these other loci should help to determine whether the various sequences that comprise the *vsp72* subfamily are mostly pseudo or functional genes, whether both types of loci are organised within tandem gene arrays, and how many functional genes are actually present within the subfamily. The underlying nature of the duplications that gave rise to these gene arrays, i.e. whether they are intrachromosomal gene duplications or partial chromosomal duplications (Adam, 1992; Upcroft *et al.* 1993a; Le Blancq & Adam, 1998), is also of fundamental interest. Determining the chromosomal location(s) of the various sequences will therefore provide important information. There is abundant evidence in the published literature, from cross- and pulse-field electrophoretic techniques, that chromosomal rearrangement events are

frequent in *G. intestinalis* within the telomeric rRNA repeat regions (Adam, 1992; Le Blancq *et al.* 1991a,b, 1992; Le Blancq, 1994; Le Blancq & Adam, 1998) and some VSP-encoding genes (Le Blancq *et al.* 1992). Le Blancq *et al.* (1991, 1992) estimated a 1% mutation rate per cell division for rRNA-encoding genes. Whether the frequency of such chromosomal recombinations contributes to the expression of pseudo genes in *Giardia* remains unknown.

From the current data set, it is difficult to envisage clearly how these pseudo gene arrays might have emerged and whether they are reservoirs of sequences that undergo occasional reactivation (by intergenic recombination) to be re-expressed as new functional hybrid genes (like in trypanosomes) or, alternatively, whether they represent an obsolete (decaying) subset of what were once functional vsp genes. In the latter case, the vsp pseudogenes would be 'old' pseudo genes and they would have accumulated mutations without selection against sites that formerly (in the previously functional gene) were non-synonymous. There are, however, some observations that would seem at odds with this interpretation:

a) The maintenance in each case, of a long, apparently uncorrupted open reading frame.

b) The conservation of cysteine tetramer motifs in the putative amino acid sequence.

c) The conservation in each case of the segment encoding the trans-membrane domain and the polyadenylation signal sequence.

As is evident from the work described in this thesis and from published studies, the *Giardia* genome database is already a valuable resource. Even in its current fragmentary form, comprising a huge number of unedited single-run sequences, it has enabled investigators in different laboratories to identify extended sequences that overlap shorter genomic segments of interest and to design primers for use in amplifying

segments of unique loci, e.g. encoding key enzymes, etc., that could otherwise not be easily obtained. Although the sequences are gradually being compiled, it remains to be seen whether this is achievable for the many similar or nearly identical segments that belong to the vsp gene family. Indeed, for loci such as *vsp*136-2 or *vsp*136-4 which contain 23.5 and 20.5 copies respectively of the same repeat unit, it may prove difficult even from 1100-bp uncorrected genome sequence runs to determine how many tandem repeat units are associated with particular sequence runs and therefore which loci these sequences represent. This problem is compounded further for loci that possess similar or identical flanking sequences, e.g. the *vsp*136 loci. For these reasons, it was considered important in this project to physically isolate (clone), map and characterise the various genomic fragments that were chosen for examination, so that the sequences could be properly distinguished and eventually compiled.

Pseudo genes have been found to have an important role in adding to the diversity of the VSG repertoire that can be expressed by *Trypanosoma equiperdum*. During the course of an infection, an extensive repertoire of VSG is expressed sequentially in African trypanosomes. There is evidence that recombination events generate diversity by re-assorting sequences and that they also allow the expression of pseudogenes (Thon *et al.* 1989; Roth *et al.* 1986; Longacre & Erisen, 1986). Using a procedure that examines the protection of VSG mRNA by genes present in genomic DNA against digestion by nuclease S1, Longacre and Erisen (1986) determined that genes coding for VSG expressed late during infections are composite genes formed from more than one Basic Copy (BC) gene. In support of this model, Roth *et al.* (1986), showed that two of the Expression Linked Copy (ELC) genes are composite loci created from the fusion of three different pseudo genes. Thon *et al.* (1989) extended their observations to another family of VSG genes by showing that a late expressed gene (VSG-20) was also created by recombination events involving another three pseudo

genes. It has been suggested that composite genes in *Trypanosoma* provide genetic diversity in the exposed N-terminal regions of VSG (Pays *et al.* 1989).

The 5' truncated nature of the *vsp72* pseudo genes is similar to pseudo genes identified in *Neisseria*. The latter pseudo genes specify a major part of the pilin protein (the pilus subunit of hair-like structures protruding from the bacterial surface), which is essential for colonisation of the human mucosa. Antigenic variation of the pilus proteins is implicated in the immune evasion and adherence properties of these pathogens (Hass & Meyer, 1986). In *Neisseria gonorrhoeae* strain MS11, a total of five silent pilin loci were identified (Hass *et al.* 1992). The pilin loci each contain multiple pilin genes, some functional and some nonfunctional. The latter lack the 5' terminal sequence encoding the N-terminal segment of the pilin polypeptide and they are referred to as 'silent' pilin copies (Hass *et al.* 1992). Most of these silent loci contain 2, 3 or 6 silent copies which are tandemly arranged (Hass *et al.* 1992). Uni-directional recombination of silent pilin DNA into an expressed pilin gene is believed to allow for substantial sequence variation of these highly immunogenic surface structures (Wainwright *et al.* 1997). The discovery in this project of tandemly arrayed vsp pseudo genes opens a new window to possibilities of vsp gene expression in *Giardia* and whether mechanisms similar to the recombinations observed in African trypanosomes and *Neisseria* and implicated in antigenic variation in these organisms, can be extended to *Giardia*.

The *in situ* mRNA hybridisation experiments were undertaken with the hope of estimating the number of loci from each gene subfamily that can be expressed during long-term axenic culture. The result of testing anti-sense probes, corresponding to different segments of *vsp*136 and *vsp*72 subfamily loci and a segment of *vsp*417-6, seemed to indicate that the axenic culture conditions favour the growth of some variant trophozoites, expressing particular but unidentified vsp genes, over the growth of other variants. This was an important finding, which indicates that methods used to estimate the VSP repertoire using

mAb's (e.g Nash *et al.* 1990a) may not provide reliable estimates of the true repertoire. Studies on the competitive growth of variants bearing different VSP *in vivo* in different hosts (e.g. Singer *et al.* 2000) will provide interesting additional information on the biological properties of these surface proteins.

Future studies

Much remains to be learnt about the precise molecular events involved in the generation and maintenance of the VSP repertoire in *Giardia* and about the stability and expression of these genes. With respect to the *vsp72* gene subfamily, it would seem worthwhile examining the remaining cloned genomic inserts to determine how many contain pseudo gene arrangements similar to the tandem gene arrays found in the fragments described in Chapter 4. It would also be informative to obtain sequence data on other functional members of the subfamily, since this would shed light on what structural differences exist between the encoded VSP and how these proteins have evolved. A first step would be to identify whether these vsp gene arrays consist of pseudo genes only or both pseudo and functional genes. The plasmid construct 21-1 would be a useful starting point, as the genomic insert has been shown to contain what appears to be a functional *vsp*72 locus. Additionally, limited PCR analysis of this construct has indicated the presence of other *vsp*72 genes, although whether these are functional or pseudo remains to be determined. Southern analysis and complete sequence determination of the pQ1 gene (section 4.10) and its flanking sequences should help determine whether this locus is part of a tandem gene array also. Similar approaches may be applied for the other constructs that were identified in this study and found (by PCR analysis and in some cases by Southern analysis also) to contain more than one *vsp*72 gene. Comparing the organisational structure of these different gene arrays and the nature of the individual pseudo (or functional) genes may help to elucidate the way in which vsp genes generally evolve, recombine and switch expression. One aspect that would

be of particular value is the identification, by karyotypic analysis and Southern hybridisation, of the chromosomal locations of the various members of the two vsp gene subfamilies described in this thesis.

Other simple and informative experiments would be genomic Southern hybridisations, using DNA from the Ad-1, WB and possibly other Group I isolates, with probes corresponding to various regions from the two vsp families described herein. This would yield information about the degree of rearrangements within the genetically similar isolates that represent this sublineage of *G. intestinalis*.

References

Acha, P. and. Szyfres, B. (1987). Zoonoses and communicable diseases common to man and animals: Scientific Publication. Pan American Health Orgenisation, Washington *503*.

Adam, R., Nash, T.E. and Wellems, T.E. (1988a). The *Giardia lamblia* trophozoite contains sets of closely related chromosomes. Nucleic Acids Research *16*, 4555-4567.

Adam, R., Aggarwal, A., Lal, A.A., Cruz, V.F., McCutchan, T. and Nash, T.E. (1988b). Antigenic variation of a cysteine-rich protein in *Giardia lamblia*. Journal of Experimental Medicine *167*.

Adam, R. (1991). The biology of *Giardia* spp. Microbiological Reviews *55*, 706-732.

Adam, R., Nash, T.E. and Wellems, T.E. (1991). Telomeric location of *Giardia* rDNA genes. Molecular and Cellular biology *11*, 3326-3330.

Adam, R. (1992). Chromosome-size variation in *Giardia lamblia*: the role of rDNA repeats. Nucleic Acids Research *20*, 3057-3061.

Adam, R., Yang, Y.M. and Nash, T.E. (1992). The cysteine-rich protein gene family of *Giardia lamblia*: loss of the *CRP170* gene in an antigenic variant. Molecular and Cellular Biology *12*, 1194-1201.

Aggarwal, A. and Nash, T. (1987). *Giardia lamblia*: RNA translation products. Experimental Parasitology *64*, 336-341.

Aggarwal, A. and Nash, T. (1988). Antigenic variation of *Giardia lamblia in vivo*. Infection and Immunity. *56*, 1420-1423.

Aley, S. and Gillin, F. (1993). *Giardia lamblia*: post-translational processing and status of exposed cysteine residues in TSA 417, a variable surface antigen. Experimental Parasitology *77*, 295-305.

Aley, S., Zimmerman, M., Hetsko, M., Selsted, M.E. and Gillin, F.D. (1994). Killing of *Giardia* lamblia by criptidins and cationic neutrophil peptides. Infection and Immunity *62*, 5397-5403.

Alonso, R. and Peattie, D. (1992). Nucleotide sequence of a second alpha giardin gene and molecular analysis of the alpha giardin genes and transcripts in *Giardia lamblia*. Molecular and Biochemical Parasitology *50*, 95-104.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25, 3389-3402.

Anderson, J., Herndon, J.L. and Vonreyn, C.F. (1993). Endemic giardiasis in New Hampshire - a case control study of environmental risks. Journal of Infectious Diseases *167*, 1391-1395.

Andrews, R., Adams, M., Boreham, P.F.L., Mayrhofer, G. and Meloni, B.P. (1989). *Giardia intestinalis*: Electrophoretic evidence for a species complex. International Journal for Parasitology *19*, 183-190.

Andrews, R., Chilton, N.B. and Mayrhofer, G. (1992). Selection of specific genotypes of *Giardia intestinalis* by growth *in vitro* and *in vivo*. Parasitology *105*, 375-386.

Andrews, R., Chilton, N.B., Ey, P.L. and Mayrhofer, G. (1993). Additional enzymes for the genetic characterization of *Giardia* from different host species. Parasitology Research *79*, 337-339.

Bangs, J., Ransom, D.M., McDowell, M.A. and Brouch, E.M. (1997). Expression of bloodstream variant surface glycoproteins in procyclic stage *Trypanosoma brucei*: role of GPI anchors in secretion. The EMBO Journal *16*, 4285-4294.

Boothroyd, J., Wang, A., Campbell, D.A. and Wang, C.C. (1987). An unusually compact ribosomal DNA repeat in the protozoon *Giardia lamblia*. Nucleic Acids Research *15*, 4065-4084.

Baruch, A.C., Issac-Renton, J. and Adam, R.D. (1996). The molecular epidemiology of *Giardia* lamblia: a sequence-based approach. Journal of Infectious Diseases *174*, 233-236.

Bernander, R., Palm, J.E.D. and Svärd, S.G. (2001). Genome ploidy in different stages of the *Giardia lamblia* life cycle. Cellular Microbiology *3*, 55-62.

Binz, N., Thompson, R.C.A., Lymbery, A.J. and Hobbs, R.P. (1992). Comparative studies on the growth dynamics of two genetically distinct isolates of *Giardia duodenalis* in vitro. International Journal for Parasitology *22*, 195-202.

Bouvier, J., Bordier, C., Vogel, H., Reichelt, R. and Etges, R. (1989). Characterization of the promastigote surface protease of *Leishmania* as a membrane-bound zinc endopeptidase. Molecular and Biochemical Parasitology *37*, 235-246.

Brodsky, W., Spencer, Jr. and Schultz, MG. (1974). Giardiasis in American travellers to the Soviet Union. Journal of Infectious Diseases *130*, 319-323.

Bruderer, T., Papanastasiou, P., Castro, R. and Konhler, P. (1993). Variant Cysteine-rich Surface Proteins of *Giardia* Isolates from Human and Animal Sources. Infection and Immunity *61*, 2937-2944.

Burnet, A., den Hollander, N., Wallis, PM., Befus, D. and Olson, ME. (1990). Zoonotic potential of giardiasis in domestic ruminants. Journal of Infectious Diseases *162*, 231-237.

Campbell, S., Van Keulen, H., Erlandsen, SL., Senturia, JB. and Jarroll, EL. (1990). *Giardia sp*: comparsion of electerophoretic karyotypes. Experimental Parasitology *71*, 470-482.

Carnaby, S., Butcher, PD., Summerbell, CD., Naeem, A. and Farthing, MJG. (1995). Minisatellites corresponding to the human polycore probes 33.6 and 33.15 in the genome of the most 'primitive' known eukaryote *Giardia lamblia*. Gene *166*, 167-172.

Chaudhuri, P.P., Sengupta, K., Manna, B., Saha, M.K. and Das, P. (1992). Detection of specific anti-*Giardia* antibodies in the serodiagnosis of symptomatic giardiasis. Journal of Diarrhoeal Disease Research *10,* 151-155.

Chen, N., Upcroft, JA. and Upcroft, P. (1994). Physical map of a 2Mb chromosome of the intestinal protozoon parasite *Giardia duodenalis*. Chromosome Research *2*, 307-313.

Chen, C., and. Sarnow, P. (1995). Initiation of protein synthesis by the eukaryotic translational apparatus on circular RNAs. Science *268*, 415-417.

Chen, N., Upcroft, JA. and Upcroft P. (1995). A *Giardia duodenalis* gene encoding a protein with multiple repeats of a toxin homologue. Parasitology *111*, 423-431.

Chen, N., Upcroft, JA. And Upcroft, P. (1996). A new cysteine-rich protein-encoding gene family in *Giardia duodenalis*. Gene *169*, 33-38.

Clark, J. and Holberton, D. (1986). Plasma membrane isolated from *Giardia lamblia*: identification of membrane proteins. European Journal of Cell Biology *42*, 200-206.

Coleman, J. (1992). Zinc proteins: enzymes, storage proteins, transcription factors, and replication proteins. Annual Review of Biochemistry *61*, 897-946.

Das, S., Schteingart, CD., Hofmann, AF., Reiner, DS., Aley, SB. and Gillin, FD. (1997). *Giardia lamblia*: Evidence for carrier-mediated uptake and release of conjugated bile acids. Experimental Parasitology *87*, 133-141.

Das, S., Traynor-Kaplan, A., Reiner, DS., Meng, TC. and Gillin, FD. (1991). A surface antigen of *Giardia lamblia* with a glycosylphosphosphatidylinositol anchor. The American Society for Biochemistry and Molecular Biology *266*, 21318-21325.

Davies, R.B. and Hibler, C.P. (1979). Animal reservoirs and cross-species transmission of *Giardia*. Waterborne transmission of Giardiasis (ed. Jakubowski,W & Hoff, J.C.) U.S. Environ. Protection Agency, Cinn OH. EPA-600/9-79-001. pp. 104-126.

Dobell, C. (1932). Antony van Leewenhoek and his "little animals". Dover Publications, New York, p. 224.

Edlind, T. and Chakreborty, P. (1987). Unusual ribosomal RNA of the intestinal parasite *Giardia lamblia.* Nucleic Acids Research *15*, 7889-78901.

Edson, C.M., Farthing, M.J.G., Thorley-Lawson, D.A. and Keusch, G.T. (1986). An 88,000 Mr *Giardia lamblia* surface protein which is immunogenic in humans. Infection and Immunity *54,* 621-625.

Ehlers, M. and Riordan, J.F. (1991). Membrane proteins with soluble counterparts: role of proteolysis in the release of transmembrane proteins. Biochemistry *30*, 10065-10071.

Einfeld, D.A. and Stibbs, H.H. (1984). Identification and characterization of a major surface antigen of *Giardia lamblia*. Infection and Immunity *46,* 377-383.

Erlandsen, S.L. (1994). Biotic transmission - Is Giardiasis a zoonosis? *In:* Thompson, R.C.A., Reynoldson, J.A. & Lymbery, A.J. (eds.), *Giardia*: From Molecules to Disease. CAB International, Wallingford, U.K. pp. 83-97.

Erlandsen, S. and Bemrick, W. (1987). SEM evidence for a new species, *Giardia psittaci*. Journal of Parasitology *73*, 623-629.

Erlandsen, S., Bemrick, WJ., Wells, CL., Feely, DE., Knudson, L., Campbell, SR., Keulen, H. and Jarroll, EL. (1990). Axenic culture and characterization of *Giardia ardeae* from the Great Blue Heron(*Ardeae herodias*). Journal of Parasitology *76*, 717-724.

Erlandsen, S.L., Sherlock, L.A., Januschka, M., Schupp, D.G., Schaefer, F.W., Jakubowski, W. and Bemrick, W.J. (1988). Cross-species transmission of *Giardia* spp.: inoculation of beavers and muskrats with cysts of human, beaver, mouse, and muskrat origin. Applied and Environmental Microbiology *54*, 2777-2785.

Erlandsen, S. and Rasch, EM. (1994). The DNA Content of Trophozoites and Cysts of *Giardia lamblia* by Microdensitometric Quantitation of Feulgen Staining and Examination by Laser Scanning Confocal Microscopy. Journal of Histochemistry and Cytochemistry *42*, 1413-1426.

Ey, P., Khanna, K., Andrews, RH., Manning, PA. and Mayrhofer, G. (1992). Distinct genetic groups of *Giardia intestinalis* distinguished by restriction fragment length polymorphisms. Journal of General Microbiology *138*, 2629-2637.

Ey, P., Andrews, RH. and Mayrhofer, G. (1993a). Differentiation of major genotypes of *Giardia intestinalis* by polymerase chain reaction analysis of a gene encoding a trophozoite surface antigen. Parasitology *106*, 347-356.

Ey, P., Darby, JM., Andrews, RH. and Mayrhofer, G. (1993b). *Giardia intestinalis*: Detection of major genotypes by restriction analysis of gene amplification products. International Journal of Parasitology *23*, 591-600.

Ey, P., Khanna, K., Manning, PA. and Mayrhofer, G. (1993c). A gene encoding a 69-kilodalton major surface protein of *Giardia intestinalis* trophozoites. Molecular and Biochemical Parasitology *58*, 247-258.

Ey, PL. and Mayrhofer, G. (1993). Two genes encoding homologous 70-kDa surface proteins are present within individual trophozoites of the binucleate protozoon parasite *Giardia intestinalis*. Gene *129*, 257-262.

Ey, P., Burderer, T., Wehrli, C. and Kohler, P. (1996). Comparison of genetic groups determined by molecular and immunological analyses of *Giardia* isolated from animals and humans in Switzerland and Australia. Parasitology Research *82*, 52-60.

Ey, P.L., Mansouri, M., Kulda, J., Nohynkova, E., Monis, P.T., Andrews, R.H. and Mayrhofer, G. (1997). Genetic analysis of *Giardia* from hoofed farm animals reveals artiodactyl-specific and potentially zoonotic genotypes. Journal of Eukaryotic Microbiology *44*, 626-635.

Ey, P.L., Darby, J.M. and Mayrhofer, G. (1998). Comparison of *tsa*417-like variant-specific surface protein (VSP) genes in *Giardia intestinalis* and identification of a novel locus in genetic group II isolates. Parasitology *117*, 445-455.

Ey, P.L. and Darby, J.M. (1998). *Giardia intestinalis*: conservation of the variant-specific surface protein VSP417-1 (TSA417) and identification of a divergent homologue ended at a duplicated locus in genetic group II isolates. Experimental Parasitology *90*, 250-261.

Ey, P.L., Darby, J.M. and Mayrhofer, G. (1999). A new locus (*vsp*417-4) belonging to the *tsa*417-like subfamily of variant-specific surface protein genes in *Giardia intestinalis*. Molecular and Biochemical Parasitology *99*, 55-68.

Fan, J., Korman, S.H., Cantor, C.R. and Smith, C.L. (1991). *Giardia lamblia*: haploid genome size determined by pulsed field gel electrophoresis is less than 12Mb. Nucleic Acids Research *19*, 1905-1908.

Farthing, M.J. (1990). Immunopathology of giardiasis. Springer Seminars in Immunopathology. *12*, 269-282.

Farthing, M.J.G. (1992). *Giardia* comes of age - progress in epidemiology, immunology and chemotherapy. Journal of Antimicrobial Chemotherapy *30*, 563-565.

Farthing, M. (1994). Giardiasis as a disease. In *Giardia*: From Molecules to Disease, R. C. A. Thompson, J. A. Reynoldson and A. J. Lymbery, eds. (Wallingford: CAB International), pp. 15-37.

Farthing, M. (1997). The molecular pathogenesis of giardiasis. Journal of Pediatric Gasteroenterology and Nutrition *24*, 79-88.

Faubert, G.M., Belosevic, M., Walker, T.S., MacLean, J.D. and Meerovitch, E. (1983). Comparative studies on the pattern of infection with *Giardia* spp. in Mongolian gerbils. Journal of Parasitology *69*, 802-805.

Faubert, G. M. (1996). The immune response to *Giardia*. Parasitology Today *12*, 140-145.

Feely, D., Erlandsen, S.L. and Chase, D.G. (1984). Structure of the trophozoite and cyst. In *Giardia* and Giardiasis, S.L. Erlandsen, E.A., ed. (New York and London: Plenum Press.), pp. 3-31.

Feely, D.E. (1988). Morphology of the cyst of *Giardia microti* by light and electron microscopy. Journal of Protozoology *35*, 52-54.

Feely, D., Holberton, D.V. and Erlandsen, S.L. (1990). The biology of *Giardia*. In Giardiasis, E.A.Meyer, ed. (Amsterdam: Elsevier), pp. 11-50.

Filice, F.P. (1952). Studies on the cytology and life history of a *Giardia* from the laboratory rat. University of California Publications in Zoology *57*, 53-146.

Freemont, P.S., Hanson, I.M. and Trowsdale, J. (1991). A novel cysteine-rich sequence motif. Cell *64*, 483-484.

Gardner, M.J., Tettelin, H., Carucci, D.J., Cummings, L.M., Smith, H.O., Fraser, C.M., Venter, J.C., Hoffman, S.L. (1999). The malaria genome sequencing project: complete sequence of *Plasmodium falciparum* chromosome 2. Parassitologia. *41*, 69-75.

Garlapati, S., Chou, J. and Wang, C.C. (2001). Specific secondary structures in the capsid-coding region of Giardiavirus transcript are required for its translation in *Giardia lamblia*. Journal of Molecular Biology *308*, 623-638.

Gillin, F.D., Reiner, D.S., Gault, M.J., Douglas, H., Das, S., Wunderlich, A. & Sauch, J.F. (1987). Encystation and expression of cyst antigens by *Giardia lamblia in vitro*. Science *235*, 1040-1043.

Gillin, F.D., Boucher, S.E., Rossi, S.S. and Reiner, D.S. (1989). *Giardia lamblia*: the roles of bile, lactic acid and pH in the completion of the life cycle in vitro. Experimental Parasitology *69*, 164-174.

Gillin, F.D., Hagblom, P., Harwood, J., Aley, S.B., McCaffery, M., Reiner, D.S., So, M. and Guiney, D.G. (1990). Isolation and expression of the gene for a major surface protein of *Giardia lamblia*. Proceedings of the National Academy of Sciences USA *87*, 4463-4467.

Gillin, F.D., Reiner, D.S. and McCaffery, JM. (1991). Organelles of protein transport in *Giardia lamblia*. Parasitology Today *7*, 113-116.

Gillin, F.D., McCaffery, J.M. and Reiner, D.S. (1996). Cell Biology of the primitive eukaryote *Giardia lamblia*. Annual Review of Microbiology *50*, 679-705.

Gottstein, B., Harriman, G.R., Conrad, T.J. and Nash, T.E. (1990). Antigenic variation in *Giardia lamblia*: cellular and humoral immune response in a mouse model. Parasite Immunology *12*, 659-673.

Gottstein, B. and Nash, T.E. (1991). Antigenic variation in *Giardia lamblia*: infection of congenitally athymic nude and *scid* mice. Parasite Immunology *13*, 649-659.

Gottstein, B., Deplazes, P. and Tanner, I. (1993). *In vitro* synthesized immunoglobulin A from nu/+ and reconstituted nu/nu mice against a dominant surface antigen of *Giardia lamblia*. Parasitology Research *79*, 644-648.

Goltz, J.P. (1980). Giardiasis in chinchillas. MS thesis, University of Guelph, Guelph, Ontario, Canada .

Grimmond, T.R., Radford, A.J., Brownridge, T., Farshid, A., Harris, C., Turton, P. and Wordsworth, K. (1988). *Giardia* carriage in aboriginal and non-aboriginal children attending urban day-care centres in South Australia. Australian Paediatric Journal *24*, 304-305.

Hare, D., Jarroll, E.L. and Lindmark, D.G. (1989). *Giardia lamblia*: Characterization of proteinase activity in trophozoites. Experimental Parasitology *68*, 168-175.

Hashimoto, T., Nakamura, Y., Shirakura, T., Adachi, J., Goto, N, Okamoto, K. and Hasegawa, M. (1994). Protein phylogeny gives a robust estimation for early divergences of eukaryotes: Phylogenetic place of a mitochondria-lacking protozoon, *Giardia lamblia*. Molecular Biology and Evolution *11*, 65-71.

Hashimoto, T., Nakamura, Y., Kamaishi, T., Nakamura, F., Adachi, J., Okamoto, K.-i. & Hasegawa, M. (1995). Phylogenetic place of mitochondrion-lacking protozoan, *Giardia lamblia*, inferred from amino acid sequences of elongation factor 2. Molecular Biology and Evolution *12*, 782-793.

Hass, R. and Meyer, T. (1986). The repertoire of silent pilus genes in *Neisseria gonorrhoeae*: evidence for gene conversion. Cell *44*, 107-115.

Hass, R., Veit, S. and Meyer, T.F. (1992). Silent pilin genes of *Neisseria gonorrhoeae* MS11 and the occurence of related hypervariant sequences among other gonococcal isolates. Molecular Microbiology *6*, 197-208.

Healey, A., Mitchell, R., Upcroft, J.A., Boreham, P.F.L. and Upcroft, P. (1990). Complete nucleotide sequence of the ribosomal RNA tandem repeat unit from *Giardia intestinalis*. Nucleic Acids Research *18*, 4006.

Heyworth, M.F., Carlson, J.R. and Ermak, T.H. (1987). Clearance of *Giardia muris* infection requires helper/inducer T lymphocytes. Journal of Experimental Medicine *165*,: 1743-1748.

Holberton, D. and Marshall, J. (1995). Analysis of consensus sequence patterns in *Giardia* cytoskeleton gene promoters. Nucleic Acids Research *15*, 2945-2953.

Homan, W., Van Enckevort, F.H.J., Limper, L., Van Eys, G.J.J.M., Schoone, G.J., Kasprzak, W., Majewska, A.C. and van Knapen, F. (1992). Comparison of *Giardia* isolates from different laboratories by isoenzyme analysis and recombinant DNA probes. Parasitology Research *78*, 316-323.

Hou, G., Le Blancq, S.M., E, Y., Huixiang, Z. and Lee, M.G.S. (1995). Structure of a frequently rearranged rRNA-encoding chromosome in *Giardia lamblia*. Nucleic Acids Research *23*, 3310-3317.

Hughes, M. and Andrews, D. (1997). A single nucleotide is a sufficient 5` untranslated region for translation in an eukaryotic *in vitro* system. FEBS Letters *414*, 19-22.

Isaac-Renton, J.L., Cordeiro, C., Sarafis, K. and Shahriari, H. (1993). Characterization of *Giardia duodenalis* isolates from a waterborne outbreak. Journal of Infectious Diseases *167*, 431-440.

Ish-Horowicz, D. and Burke, J.F. (1981). Rapid and efficient cosmid cloning. Nucleic Acids Research *9*, 2989-2998.

Jakubowski, W. (1990). The control of *Giardia* in water supplies. In Giardiasis, E. A. Meyer, ed. (Amsterdam: Elsevier), pp. 335-353.

Kabnick, S. and Peattie, D. (1990). *In situ* analyses reveals that two nuclei of *Giardia lamblia* are equivalent. Journal of Cell Science *95*, 353-360.

Kappus, K.D., Lundgren, R.G. Jr., Juranek, D.D., Roberts, J.M. and Spencer, H.C. (1994). Intestinal parasitism in the United States: update on a continuing problem. American Journal of Tropical Medicine and Hygiene *50*, 705-713.

Karanis, P. and Ey, P.L. (1998). Characterization of axenic isolates of *Giardia intestinalis* established from humans and animals in Germany. Parasitology Research *84*, 442-449.

Karapetyan, A.E. (1960). Methods of *Lamblia* cultivation. Tsitologiia *2*, 379-384.

Kattenbach, W, Pimenta, P.F.P, Souza, de W. and Pinto da Silva, P. (1991). *Giardia duodenalis*: a freeze-fracture, fracture-flip and cytochemistry study. Parasitology Research *77*, 651-658.

Keister, D. (1983). Axenic culture of *Giardia lamblia* in TYI-S-33 medium supplememted with bile. Transactions of the Royal Society of Tropical Medicine and Hygiene *77*, 487-488.

Korman, S., Le Blancq, S.M., Deckelbaum, R.J. and van der Ploeg, L.H.T. (1992). Investigation of human giardiasis by karyotype analysis. Journal of Clinical Investigation *89*, 1725-1733.

Kirk-Mason, K., Turner, M.J. and Chakraborty, P.R. (1989). Evidence for unusually short tubulin mRNA leaders and characterisation of tubulin genes in *Giardia lamblia*. Molecular and Biochemical Parasitology *36*, 87-100.

Knodler, L.A., Svard, S.G., Silberman, J.D., Davids, B.J. and Gillin, F.D. (1999). Developmental gene regulation in *Giardia lamblia*: first evidence for an encystation-specific promoter and differential 5' mRNA processing. Molecular Microbiology *34*, 327-340.

Kulda, J. and Nohýnková, E. (1996). *Giardia* in humans and animals. In: "Parasitic Protozoa", 2nd edn. (ed. P.P. Kreier), Academic Press, San Diego 10: 225-422.

Kumar, S., Tamura, K. and Nei M. (1994). MEGA - Molecular Evolutionary Genetics Analysis software for microcomputers. Computers in the Applied Biosciences *10*, 189-191.

Kumkum, R., Khanna, M., Mehta, S. and Vinayak, VK. (1988). Plasma membrane associated antigens of trophozoites of axenic *Giardia lamblia*. The Royal Society of Tropical Medicine and Hygiene. *82*, 439-444.

Laemmli, U. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature *227*, 680-685.

Lanfredi-Rangel, A., Kattenbach, W.M., Diniz, J.A. Jr., de Souza, W. (1999). Trophozoites of *Giardia lamblia* may have a Golgi-like structure. FEMS Microbiol Lett. 181, 245-51.

Laurent, M., Pays, E., van der Werf, A., Aerts, D., Magnus, E., van Meirvenne, N. and Steinnert, M. (1984). Nucleic Acids Research *12*, 8319-8328.

Le Blancq, S., Kase, SR. and van der Ploeg, L.H.T. (1991a). Analysis of a *Giardia lamblia* rRNA encoding telomere with (TAGGG)n as the telomere repeat. Nucleic Acids Research *19*, 5790.

Le Blancq, S., Korman, S.H. and van der Ploeg, L.H.T. (1991b). Frequent rearrangements of rRNA-encoding chromosomes in *Giardia lamblia*. Nucleic Acids Research *19*, 4405-4412.

Le Blancq, S., Korman, S.H. and van der Ploeg, L.H.T. (1992). Spontaneous chromosome rearrangement in the protozoon *Giardia lamblia*: estimation of mutation rates. Nucleic Acids Reseach *20*, 4539-4545.

Le Blancq, S. (1994). Chromosome rearrangements in *Giardia lamblia*. Parasitology Today *10*, 177-179.

Le Blancq, S. and Adam, R. (1998). Structural basis of karyotype heterogenecity in *Giardia lamblia*. Molecular and Biochemical Parasitology *97*, 199-208.

LeChevallier, M.W., Norton, W.D. and Lee, R.G. (1991). Occurrence of Giardia and Cryptosporidium spp. in surface water supplies. Applied and Environtal Microbiology *57*, 2610-6.

Longacre, S. and Eisen, H. (1986). Expression of whole and hybrid genes in *Trypanosoma equiperdum* antigenic variatiion. The EMBO Journal *5*, 1057-1063.

Lu, S.O., Branch, A.C. and Adam, R.D. (1998). Molecular comparison of *Giardia* lamblia isolates. Journal of Parasitology *28*, 1341-1345

Lujan, H., Mowatt, M.R., Wu, J., Lu, Y., Lees, A., Chance, M.R. and Nash, T.E. (1995). Purification of a variant-specific surface protein of *Giardia lamblia* and characterization of its metal-binding properties. Journal of Biological Chemistry *270*, 13807-13813.

Lujan, H.D., Mowatt, M.R. and Nash, T.E. (1998). The molecular mechanisms of *Giardia* encystation. Parasitology Today *14*, 446-450.

Lunn, P., Erinoso, H.O., Northrop-Clewes, C.A. and Boyce, S.A. (1999). *Giardia intestinalis* is unlikely to be a major cause of the poor growth of rural Gambian infants. Journal of Nutrition *129*, 872-877.

McArthur, A.G., Morrison, H.G., Nixon, J.E., Passamaneck, N.Q., Kim, U., Hinkle, G., Crocker, M.K., Holder, M.E., Farr, R., Reich, C.I., Olsen, G.E., Aley, S.B., Adam, R.D., Gillin, F.D. and Sogin, M.L. (2000). The *Giardia* genome project database. FEMS Microbiol. Letters *189*, 271-273.

McCaffery, J.M. and Gillin, F.D. (1994). *Giardia lamblia*: ultrastructural basis of protein transport during growth and encystation. Experimental Parasitology *79*, 220-235.

Macechko, P.T., van Keulen, H., Jarroll, E.L., Mulgrew, T., Gurien, A. and Erlandsen, S.L. (1998). Detection of *Giardia* trophozoites in archival pathology specimens of human small intestine. Microscopy & Microanalysis *4*, 397-403.

Mata, L. (1978). The childern of Santa Maria Cauque: a prospective field study of health and growth. MIT Press, Cambridge, Mass.

Mayrhofer, G. and Waight-Sharma, A. (1988). The secretory immune response in rats infected with rodent *Giardia duodenalis* isolates and evidence for passive protection with immune bile. In Advances in *Giardia* Research, P.M. Wallis and B.R. Hammond, ed. (Calgary: University of Calgary Press.), pp. 49-54.

Mayrhofer, G., Andrews, R.H., Ey, P.L., Albert, M.J., Grimmond, T.R. and Merry, D.J. (1992). The use of suckling mice to isolate and grow *Giardia* from mammalian faecal specimens for genetic analysis. Parasitology *105*, 255-263.

Mayrhofer, G., Andrews, R.H., Ey, P.L. and Chilton, N.B. (1995). Division of *Giardia* isolates from humans into two genetically distinct assemblages by electrophoretic analysis of enzymes encoded at 27 loci and comparison with *Giardia muris*. Parasitology *111*, 11-17.

Meloni, B.P. and Thompson, R.C.A. (1987). Comparative studies on the axenic in vitro cultivation of *Giardia* of human and canine origin: evidence for intraspecific variation. Trans. R. Soc. Trop. Med. Hyg. *81*, 637-640.

Meloni, B.P., Lymbery, A.J. and Thompson, R.C.A. (1988). Isoenzyme electrophoresis of 30 isolates of Giardia from humans and felines. Amer. J. Trop. Med. Hyg. *38*, 65-73.

Meloni, B., Thompson, R.C.A., Stranden, A.M., Köhler, P. and Eckert, J. (1992). Critical comarsion of *Giardia duodenalis* from Australia and Switzerland using isoenzyme electrophoresis. Acta Tropica *50*, 115-124.

Meloni, B., Lymbery, A.J. and Thompson, R.C.A. (1995). Genetic characterisation of isolates of *Giardia duodenalis* by enzyme electrophoresis: implications for reproductive biology, population structure, taxonomy, and epidemiology. J. Parasitology *8*, 368-383.

Meng, T., Hetsko, M.L. and Gillin, F.D. (1993). Antigenic switching of TSA 417, a trophozoite variable surface protein, following completion of the life cycle *of Giardia lamblia*. Infection and Immunity *61*, 5394-5397.

Meyer, E.A. (1976). *Giardia lamblia*: isolation and axenic cultivation. Experimental Parasitology *39*, 101-105.

Meyer, E.A. and Jarroll, E.L. (1980). Giardiasis. American Journal of Epidemiology *111*, 1-12.

Meyer, E. (1990). Taxonomy and Nomenclature. In Giardiasis, E. A. Meyer, ed. (Amsterdam: Elsevier), pp. 51-60.

Meyne, J., Ratliff, R.L. and Moyzis, R.K. (1989). Conservation of the human telomere sequence (TTAGGG)n among vertebrates. Proceedings of the National Academy of Sciances U.S.A. *86*, 7049-7053.

Miotti, P., Gilman, R.H., Santosham, M. and Ryder, R.W. (1986). Age-related rate of seroposivity of antibody to *Giardia lamblia* in four diverse populations. Journal of Clinical Microbiology *24*, 972-975.

Moore, G.T., Cross, W.M., McGuire, D., Mollohan, C.S., Gleason, N.N., Healy, G.R. and Newton, L.H. (1969). Epidemic giardiasis at a ski resort. New England Journal of Medicine *281*, 402-407.

Monis, P.T., Mayrhofer, G., Andrews, R.H., Homan, W.L., Limper, L. and Ey, P.L. (1996). Molecular genetic analysis of *Giardia intestinalis* isolates at the glutamate dehydrogenase locus. Parasitology *112*, 1-12.

Monis, P.T., Andrews, R.H., Mayrhofer, G., Mackrill, J., Kulda, J., Renton, J.L and Ey, P.L. (1998). Novel lineages of *Giardia intestinalis* identified by genetic analysis of organisms isolated from dogs in Australia. Parasitology *116*, 7-19.

Monis, P.T., Andrews, R.H., Mayrhofer, G. and Ey, P.L. (1999). Molecular systematics of the parasitic protozoon *Giardia intestialis*. Molecular Biology and Evolution *16*, 1135-1144.

Mowatt, M., Aggarwal, A. and Nash, T.E. (1991). Carboxy-terminal sequence conservation among variant-specific surface proteins of *Giardia lamblia*. Molecular and Biochemical Parasitology *49*, 215-228.

Mowatt, M., Nguyen, B.T., Conrad, J.T., Adam, R.D. and Nash, T.E. (1994). Size heterogeneity among antigenically related *Giardia lamblia* variant-specific surface proteins is due to differences in tandem repeat copy number. Infection and Immunity *62*, 1213-1218.

Muller, N., Stager, S. and Gottstein, B. (1996). Serological analysis of antigenic heterogeneity of *Giardia lamblia* variant surface proteins. Infection and Immunity *64*, 1385-1390.

Müller, N. and Gottstein, B. (1998). Antigenic variation and the murine immune response to *Giardia lamblia*. International Journal of Parasitology *28*, 1829-1839.

Müller, N. and Stäger, S. (1999). Periodic appearance of a predominant variant antigen type during a chronic *Giardida lamblia* infection in a mouse model. International Journal of Parasitology *29*, 1917-1923.

Myler, P.J., Allison, J., Agabian, N. and Stuart, K. (1984). Cell *39*, 203-211.

Nash, T., Gillin, F.D. and Smith, P.D. (1983). Excretory-Secretory Product of *Giardia lamblia*. Journal of Immunology *131*, 2004-2010.

Nash, T., McCutchan, T., Keister, D., Dame, J.B., Conrad, J.D. and Gillin, F.D. (1985). Restriction-endonuclease analysis of DNA from 15 *Giardia* isolates obtained from humans and animals. The Journal of Infectious Disease *152*, 64-72.

Nash, T. and Keister, D. (1985). Differences in excretory-secretory products and surface antigens among 19 isolates of *Giardia*. The Journal of Inferctious Disease *152*, 1166-1171.

Nash, T., Herrington, D.A., Losonsky, G.A. and Levine, M.M. (1987). Experimental human infections with *Giardia lamblia*. Journal of Infectious Diseases *156*, 974-984.

Nash, T.E., Aggarwal, A., Adam, R.D., Conrad, J.T. and Merritt, J.W.Jr. (1988). Antigenic variation in *Giardia lamblia*. Journal of Immunology *141*, 636-641.

Nash, T. (1989). Antigenic variation in *Giardia lamblia*. Experimental Parasitology *68*, 238-241.

Nash, T., Banks, S.M., Alling, D.W., Merritt, J.J.W. and Conrad, J.T. (1990a). Frequency of variant antigens in *Giardia lamblia*. Experimental Parasitology *71*, 415-421.

Nash, T., Conrad, J.T. and Merritt, J.J.W. (1990b). Variant specific epitopes of *Giardia lamblia*. Molecular and Biochemical Parasitology *42*, 125-132.

Nash, T., Herrington, D.A., Levine, M.M., Conrad, J.T. and Merritt, J.J.W. (1990c). Antigenic variation of *Giardia lamblia* in experimental human infections. Journal of Immunology *144*, 4362-4369.

Nash, T., Merritt, J.J.W. and Conrad, J.T. (1991). Isolate and epitope variability in susceptibility of *Giardia lamblia* to intestinal proteases. Infection and Immunity *59*, 1334-1340.

Nash, T.E. (1992). Surface antigen variability and variation in *Giardia lamblia*. Parasitology Today *8*, 229-234.

Nash, T.E. And Mowatt, M.R. (1992a). Identification and characterisation of a *Giardia lamblia* group-specific gene. Experimental Parasitology *75*, 369-378.

Nash, T.E. and Mowatt, M.R. (1992b). Characterization of a *Giardia lamblia* variant-specific surface protein (VSP) gene from isolate GS/M and estimation of the VSP gene repertoire size. Molecular and Biochemical Parasitology *51*, 219-228.

Nash, T.E. and Mowatt, M.R. (1993). Variant-specific surface proteins of *Giardia lamblia* are zinc-binding proteins. Proceedings of the National Academy of Science USA *90*, 5489-5493.

Nash, T.E., Conrad, J.T. and Mowatt, M.R. (1995). *Giardia* lamblia: Identification of a variant-specific surface protein gene family. Journal of Eukaryotic Microbiology *42*, 604-609.

Nash, T.E. and Rice, W. (1998). Efficiacies of zinc-finger- active drugs against *Giardia lamblia*. American Society for Microbiology *42*, 1488-1492.

Nielsen, H., Engelbrecht, J., Brunak, S. and Heijne, G.V. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering *10*, 1-6.

Ortega, Y. and Adam, R. (1997). *Giardia*: Overview and Update. Clinical Infectious Diseases *25*, 545-550.

Ortega-Pierres, R. Alonso, L. Cervantes and C. Montanez (1990). Characterization of two DNA populations of *Giardia lamblia*. Molecular Microbiology *4*, 1985-1991.

Oyerinde, P.O., Ogunbi, O. and Alonge, A.A. (1977). Age and sex distribution of infections with *Entamoeba histolytica* and *Giardia lamblia* in the Lagos population. International Journal of Epidemiology *6*, 231-234.

Papanastasiou, P., Hiltpold, A., Bommeli, C. and Köhler, P. (1996a). The release of variant surface protein of *Giardia* to its soluble isoform is mediated by the selective cleavage of the conserved carboxy-terminal domain. Biochemistry *35*, 10143-10148.

Papanastasiou, P., McConville, M.J., Ralton, J. and Köhler, P. (1996b). The variant-specific surface protein of *Giardia*, VSP4A1, is a glycosylated and palmitoylated protein. Journal of Biochemistry *322*, 49-56.

Papanastasiou, P., Bruderer, T., Li,Y., Bommeli, C. and Kohler, P. (1997). Primary structure and biochemical properties of a variant-specific surface protein of *Giardia*. Molecular and Biochemical Parasitology *86*, 13-27.

Pays, E. (1989). Pseudogenes, chimaeric genes and the timing of antigen variation in African trypanosomes. Trends in Genetics *5*, 389-391

Pays, E., Vanhamme, L. and Berberof, M. (1994). Genetic controls for the expression of surface antigens in African trypanosomes. Annual Reviews of Microbiology *48*, 25-52.

Peattie, D.A., Alonso, R.A., Hein, A. and Caulfield, J.P. (1989). Ultrastructural localization of giardins to the edges of disk microribbons of *Giardia lamblia* and the nucleotide and deduced protein sequence of alpha giardin. Journal of Cell Biology *109*, 2323-2335.

Pimenta, P., DaSilva, P.P. and Nash, T.E. (1991). Variant surface antigens of *Giardia lamblia* are associated with the presence of a thick cell coat: thin section and label fracture immunocytochemistry survey. Infection and Immunity *59*, 3989-3996.

Quick, R., Paugh, K., Addiss, D., Kobayashi, J. and Baron, R. (1992). Restaurant-associated outbreak of giardiasis. Journal of Infectious Diseases *166*, 673-676.

Rosales-Borjas, D., Diaz-Rivadeneyra, J., Dona-Leyva, A., Zambrano-Villa, SA., Mascaro, C., Osuna, A. and Ortiz-Ortiz, L. (1998). Secretory immune response to membrane antigens during *Giardia lamblia* infection in humans. Infection and Immunity *66*, 756-759.

Rendtorff, R.C. (1954). Experimental transmission of human intestinal protozoon parasites. IV. Attempts to transmit *Entamoeba coli* and *Giardia lamblia* cysts by water. American Journal of Hygiene *60*, 327-338.

Roth, C. W., Longacre, S., Raibaud, A., Baltz, T. and Eisen, H. (1986). The use of incomplete genes for the construction of a *Trypanosoma equiperdum* variant surface glycoprotein gene. The EMBO Journal *5*, 1065-1070.

Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989). Molecular cloning. A Laboratory Manual, 2nd edn.. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Sanchez-Garcia, I. and Rabbitts, T.H. (1994). The LIM domain: a new structural motif found in zinc-finger-like proteins. Trends in Genetics *10*, 315-20.

Sarafis, K. and Isaac-Renton, J. (1993). Pulsed-field gel electrophoresis as a method of biotyping of *Giardia duodenalis*. American Journal of Tropical Medicine and Hygiene. *48*, 134-144.

Schmererin, M., Jones, T.C. and Klein, H. (1978). Giardiasis: association with homosexuality. Annals of Internal Medicine *88*, 801-803.

Sheppard, M., Thompson, J.K., Anders, R.F., Kemp, D.J. and Lew, A.M. (1989). Molecular karyotyping of the rodent malarias *Plasmodium chabaudi*, *Plasmodium Berghei*, and *Plasmodium cinckei*. Molecular and Biochemical Parasitology *34*, 45-52.

Sibley, L. and Boothroyd, J. (1992). Construction of a molecular karyotype for *Toxoplasma gondii*. Molecular and Biochemical Parasitology *51*, 291-300.

Singer, S.M., Elmendorf, H.G., Conrad, J.T. and Nash, T.E. (2000). Biological selection of variant-specific surface proteins in *Giardia lamblia*. Journal of Infectious Diseases *16*, 183-188.

Singer, S., Yee, J. and Nash, T.E. (1998). Episomal and integrated maintenance of foreign DNA in *Giardia lamblia*. Molecular and Biochemical Parasitology *92*, 59-69.

Singer, S. and Nash, T.E. (2000a). T-cell dependent control of acute *Giardia lamblia* infections in mice. Infection and Immunity *68*, 170-175.

Singer, S. and Nash, T. (2000b). The role of normal flora in *Giardia lamblia* infections in mice. Journal of Infectious Diseases *181*, 1510-1512.

Smith, P.D., Gillin, F.D., Kaushal, N.A. and Nash, T.E. (1982). Antigenic analysis of *Giardia lamblia* from Afganistan, Puerto Rico, Ecuador, and Oregon. Infection and Immunity *36*, 714-719.

Smith, M., Aley, S.B., Sogin, M,. Gillin, F.D. and Evans, G.A. (1998). Sequence survey of the *Giardia lamblia* genome. Molecular and Biochemical Parasitology *95*, 267-280.

Snider, D., Gordon, J., McDermott, M.R. and Underdown, B.J. (1985). Chronic *Giardia muris* infection in anti-IgM-treated mice. I. Analysis of immunoglobolin and parasite-specific antibody in normal and immunoglobolin-deficient animals. Journal of Immunology *134*, 4153-4162.

Snider, D., Skea, D. and Underdown, B.J. (1988). Chronic giardiasis in B cell-deficient mice expressing the xid gene. Infection and Immunity *56*, 2838-2842.

Sogin, M., Gunderson, J.H., Elwood, H., Alonso, R.A. and Peattie, D.A. (1989). Phylogenetic Meaning of the Kingdom Concept: An Unusual Ribosomal RNA from *Giardia lamblia*. Science *243*, 75-77.

Soliman, M., Taghi-Kilani, R., Abou-Shady, A.F.A., El-Mageid, S.A.A., Handousa, A.A., Hegazi, M.M. and Belosevic, M. (1998). Comparison of serum antibody responses to *Giardia lamblia* of symptomatic and asymptomatic patients. American Journal of Tropical Medicine and Hygiene *58*, 232-239.

Stäger, S. and Muller, N. (1997). *Giardia lamblia* infections in B-cell-deficient transgenic mice. Infection and Immunity *65*, 3944-3946.

Stäger, S., Felleisen, R., Gottstein, B. and Muller, N. (1997). *Giardia lamblia* variant surface protein H7 stimulates a heterogenous repertoire of antibodies displaying differential cytological effects on the parasite. Molecular and Biochemical Parasitology *85*, 113-124.

Stäger, S., Gottstein, B., Sager, H., Jungi, W. and Muller, N. (1998). Influence of antibodies in mother's milk on antigenic variation of *Giardia lambllia* in the murine mother-offspring model of infection. Infection and Immunity *66*, 1287-1292.

Steketee, R.W., Reid, S., Cheng, T., Stoebig, J.S., Harrington, R.G. and Davis, J.P. (1989). Recurrent outbreaks of giardiasis in a child day care center, Wisconsin. American Journal of Public Health *84*, 450-451.

Stevens, D.P., Frank, D.M. and Mahmoud, A.A.F. (1978). Thymus dependency of host resistance to Giardia muris infection: studies in nude mice. Journal of Immmunology *120,* 680-682.

Sun, C., Chou, C.F. and Tai, J.H. (1998). Stable DNA transfection of the primitive protozoon pathogen *Giardia lamblia*. Molecular and Biochemical Parasitology *92*, 123-132.

Svärd, S.G, Meng T.C, Hetsko, M.L, McCaffery, J.M and Gillin, F.D. (1998). Differentiation-associated surface antigen variation in the ancient eukaryote *Giardia lamblia*. Molecular Microbiology *30*, 979-989.

Swarbrick, A., Lim, R.L.H., Upcroft, J.A. and Stewart, T.S. (1997). Nucleotide variation in the cytidine triophosphate synthetase gene of *Giardia duodenalis*. Journal of Eukaryotic Microbiology *44*, 531-534.

Thompson, R., Lymbery, A.J. and Meloni, B.P. (1990). Genetic variation in *Giardia* Kunstler. Protozoology Abstracts *14*, 1-28.

Thompson, R. and Meloni, B. (1993). Molecular variation in *Giardia*. Acta Tropica *53*, 167-184.

Thompson, S.C. (1994). *Giardia lamblia* in children and the child care setting: a review of the literature. Journal of Paediatrics and Child Health *30*, 202-209.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Research *22,* 4673-4680.

Thon, G., Baltz, T. and Harvey, E. (1989). Antigenic diversity by the recombination of pseudogenes. Genes and Development *3*, 1247-1254.

Tibayrenc, M. (1994). How many species of *Giardia* are there? In *Giardia*: From Molecules to Disease, R.C.A. Thompson, J.A. Reynoldson and A.J. Lymbery, eds. (Wallingford: CAB International), pp. 41-48.

Upcroft, J., Boreham, P.F.L. and Upcroft, P. (1989). Geographic variation in *Giardia* karyotypes. International Journal for Parasitology *19*, 519-527.

Upcroft, P. (1991). DNA Fingerprinting of the human intestinal parasite *Giardia intestinalis* with hypervariable minisatellite sequences. In *DNA Fingerprinting: Approaches and Applications*, T. Burke, Dolf,G., Jefferys,A.J. and Wolff,R., ed. (Switzerland: Birkhauser Verlag Basel), pp. 70-84.

Upcroft, J., Healey, A., Murray, D.G., Boreham, P.F.L. and Upcroft, P. (1992). A gene associated with cell division and drug resistance in *Giardia duodenalis*. Parasitology *104*, 397-405.

Upcroft, P. and Healey, A. (1993). PCR priming from the restriction endonuclease site 3' extension. Nucleic Acids Research *21,* 4854.

Upcroft, J., Healey, A. and Upcroft, P. (1993a). Chromosomal duplication in *Giardia duodenalis*. International Journal of Parasitology *23*, 609-616.

Upcroft, J., Mitchell, R.W. and Upcroft, P. (1993b). The 3` terminal region of a gene encoding a cysteine rich surface protein in *Giardia duodenalis*. International Journal for Parasitology *23*, 785-792.

Upcroft, J.A. & Upcroft, P. (1994). Two distinct varieties of *Giardia* in a mixed infection from a single patient. Journal of Eukaryotic Microbiology *41*, 189-194.

Upcroft, P., Chen, N. and Upcroft, J.A. (1997). Telomeric organization of a variable and inducible toxin gene family in the ancient eukaryote *Giardia duodenalis*. Genome Research *7*, 37-46.

Vanhamme, L. and Pays, E. (1995). Control of gene expression in Trypanosomes. Microbiological Reviews *59*, 223-240.

Van Keulen, H., Gutell, R.R., Gates, M.A., Campbell, S.R., Erlandsen, S.L., Jarroll, E.L., Kulda, J. and Meyer, E.A. (1993). Unique phylogenetic position of Diplomonidida based on the complete small subunit ribosomal RNA sequence of *Giardia ardeae*, *G. muris*, *G. duodenalis* and *Hexamita* sp.. The FASEB Journal *7*, 223-231.

Vinayak, V.K., Kum Kum., Venkateswarlu, K., Khanna, R. and Mehta, S. (1987). Hypo-gammaglobulinaemia in children with persistent giardiasis. Journal of Tropical Pediatrics *33*, 140-142.

Visvesvara, G., Dickerson, J.W. and Healy, G.R. (1988). Variable infectivity of human-derived *Giardia lamblia* cysts for Mongolian gerbils. Journal of Clinical Microbiology *26*, 837-841.

Wainwright, L.A., Frangipane, J.V. and Seifert, H.S. (1997). Analysis of protein binding to the *Sma/Cla* DNA repeat in pathogenic *Neisseriae*. Nucleic Acids Research *25*, 1362-1368.

Waight-Sharma, A. and Mayrhofer, G. (1988a). Biliary antibody response in rats infected with rodent *Giardia duodenalis* isolates. Parasite Immunology *10*, 181-191.

Waight-Sharma, A. and Mayrhofer, G. (1988b). A comparative study of infections with rodent isolates of *Giardia duodenalis* in inbred strains of rats and mice and in hypothymic nude rats. Parasite Immunology *10*, 169-179.

Wallis, P.M., Buchanan-Mappin, J.M., Faubert, G.M. and Belosevic, M. (1984). Reservoirs of *Giardia* spp. in southwestern Alberta. Journal of Wildlife Diseases *20*, 279-283.

Weber, K., Schneider, A., Westermann, S., Muller, N. and Plessmann, U. (1997). Posttranslational modifications of a- and b-tubulin in *Giardia lamblia*, an ancient eukaryote. FEBS Letters *419*, 87-91.

Weiss, J., van Keulen, H. and Nash, T.E. (1992). Classification of subgroups of *Giardia lamblia* based upon ribosomal RNA gene sequence using the polymerase chain reaction. Molecular and Biochemical Parasitology` *54*, 73-86.

Wilkens, L., Tchinda, J., Komminoth, P. and Werner, M. (1997). Single- and double-color oligonucleotide primed in situ labeling (PRINS): applications in pathology. Histochem. Cell. Biol. *108*, 439-446.

Wolfe, M. (1992). Giardiasis. Clinical Microbiology Reviews *5*, 93-100.

Yang, Y. and Adam, R.D. (1994). Allele-specific expression of a variant-specific surface protein (VSP) of *Giardia lamblia*. Nucleic Acids Research *22*, 2102-2108.

Yang, Y., Ortega, Y., Sterling, C. and Adam, R.D. (1994). *Giardia lamblia* trophozoites contain multiple alleles of a variant-specific surface protein gene with 105-base pair tandem repeats. Molecular and Biochemical Parasitology *68*, 257-276.

Yang, Y. and Adam, R.D. (1995). A group of *Giardia lamblia* variant-specific surface protein (VSP) genes with nearly identical 5' regions. Molecular and Biochemical Parasitology *75*, 69-74.

Yee, J., Mowatt, M.R., Dennis, P.P. and Nash, T.E. (2000). Transcriptional analysis of the glutamate dehydrogenase gene in the primitive eukaryote, *Giardia lamblia*. Journal of Biological Chemistry *275*, 11432-11439.

Yu, D., Wang, A.L., Botka, C.W. and Wang, C.C. (1998). Protein synthesis in *Giardia lamblia* may involve interaction between a downstream box (DB) in mRNA and an anti-DB in the 16S-like ribosomal RNA. Molecular and Biochemical Parasitology *96*, 151-165

Zhang, Y., Aley, S.B., Stanley, J.S.L. and Gillin, F.D. (1993). Cysteine-dependent zinc binding by membrane proteins of *Giardia lamblia*. Infection and Immunity *61*, 520-524.

Ziegelbauer, K., Stahl, B., Karas, M., Stierhof, Y-D. and Overath, P. (1993). Proteolytic release of cell surface proteins during differentiation of *Trypanosoma brucei*. Biochemistry *32*, 3737-3742.

Zomerdijk, J., Kieft, R. and Borst, P. (1992). A ribosomal RNA gene promoter at he telomere of mini-chromosme in *Trypanosoma brucei*. Nucleic Acids Research *20*, 2725-2734.