Identification and molecular cloning of four cysteine proteinase genes from the pathogenic protozoon *Trichomonas vaginalis*

David J. Mallinson,¹† Barbara C. Lockwood,²§ Graham H. Coombs² and Michael J. North¹

Author for correspondence: Michael J. North. Tel: +44 786 467764. Fax: +44 786 464994.

¹ Department of Biological and Molecular Sciences, University of Stirling, Stirling FK9 4LA, UK

² Laboratory for Biochemical Parasitology, Department of Zoology, University of Glasgow, Glasgow G12 8QQ, UK

The parasitic protozoon *Trichomonas vaginalis* produces multiple forms of cysteine proteinase (CP). The molecular basis for this has now been examined by cloning DNA fragments encoding CPs. Using generic degenerate oligonucleotide primers based on two well-conserved regions within the central region of all eukaryotic CPs, several polymerase chain reaction fragments were isolated from *T. vaginalis* genomic DNA and shown to encode different CPs. One fragment with a well-represented sequence was used as a general probe to screen a *T. vaginalis* cDNA library at moderate stringency and five different cDNA clones were isolated. Preliminary sequencing showed that they encoded similar but distinct CPs. In the process of confirming the 5' end of one of these cDNA clones using RACE-PCR (rapid amplification of cDNA 5' ends-polymerase chain reaction), an additional sequence encoding a different CP was identified. The corresponding clone (TvCP3) and the three longest clones from the library screen (TvCP1, TvCP2 and TvCP4) were characterized further. TvCP1 and TvCP2 were full-length and TvCP3 and TvCP4 were apparently slightly less than full-length. Comparison of the predicted amino acid sequences of the four clones showed that TvCP1 and TvCP4 are related (72% identity). TvCP2 is closer to TvCP1 (60%) and TvCP4 (65%) than is TvCP3, which has 53%, 59% and 56% identity to TvCP1, TvCP2 and TvCP4, respectively. Comparison with the sequences of other known CPs indicated that the *T. vaginalis* gene products all belong to the cathepsin L/cathepsin H/papain branch of the papain superfamily. The TvCP1, TvCP2 and TvCP4 sequences are related (38–45% identity) to those of CP2 of *Dictyostelium discoideum*, human cathepsin L, three CPs from lobster and CPs from black gram, oilseed rape and rice (oryzains α and β). TvCP3 shows less identity to the other eukaryotic CPs but is most similar to *D. discoideum* CP2 (38%). The four predicted amino acid sequences share some features distinct from the majority of CPs, which suggests they might have had a common evolutionary origin. The most striking feature of sequences TvCP1, TvCP2 and TvCP3 is the apparent lack of a pre-sequence (signal sequence) for TvCP1 and very short pre-sequences for TvCP2 and TvCP3. Southern analysis indicated that the organization of the genes corresponding to the TvCP cDNAs differs. The TvCP1, TvCP2 and TvCP3 genes are single-copy, whereas the TvCP4 gene appeared to be multiple-copy. Similarly sized, single abundant transcripts were present for all four sequences. Overall, the data show that we have identified a family of genes in *T. vaginalis* which encode a number of CPs. In total, seven distinct sequences have been recognized. This suggests that the multiplicity of CP activities seen

† Present address: Beatson Institute for Cancer Research, CRC Beatson Laboratories, Garscube Estate, Switchback Road, Bearsden, Glasgow G61 1BD, UK.

§ Died 8 October 1993.

Abbreviations: RACE, rapid amplification of cDNA 5' ends; CP, cysteine proteinase; ss cDNA, single-stranded cDNA; HCL, human cathepsin L.

The nucleotide sequences of cDNAs TvCP1, TvCP2 and TvCP3 and TvCP4 have been submitted to EMBL and have been assigned the accession numbers X77218, X77219, X77220 and X77221, respectively.

in this organism is likely to be due, in part at least, to the presence of multiple genes.

Keywords: *Trichomonas vaginalis*, cysteine proteinase, papain superfamily, molecular cloning, signal peptide

INTRODUCTION

Trichomonas vaginalis is a flagellate protozoon which parasitizes the urogenital tract of humans. It is responsible for the most prevalent non-viral sexually transmitted disease, and world-wide approximately 180 million cases occur annually (Heine & McGregor, 1993). It is generally considered asymptomatic in men, although it is also known to be a cause of male urethritis (Kreiger *et al.*, 1993). In women, *T. vaginalis* causes vaginitis and exocervicitis and has recently been implicated in more serious complications such as the pathogenesis of preterm labour, premature rupture of membranes and upper reproductive tract postsurgical infections (Heine & McGregor, 1993). The pathogenesis of these disease processes is unknown. There is an effective chemotherapy for trichomoniasis using 5-nitroimidazoles, in particular metronidazole. Drug resistance, however, is now well documented, although to date this has been relatively rare and has been overcome by increased drug dosages or prolonged treatment (Heine & McGregor, 1993; Johnson, 1993). Nevertheless, there is a need for alternative strategies for chemotherapy and for the development of new drugs. As with other parasitic protozoa, cysteine proteinases (CPs) may offer suitable targets for chemotherapeutic attack (North *et al.*, 1990a; McKerrow *et al.*, 1993).

T. vaginalis contains high proteolytic activity, due almost exclusively to lysosomal CPs (North, 1991). The enzymes may be of importance in the host-parasite relationship and could contribute to pathogenicity (North, 1991). One-dimensional gelatin-SDS-PAGE analysis of cell lysates has shown that there are at least 11 different proteinase activities, all of which were shown, by the use of selective proteinase inhibitors, to be CPs (Lockwood *et al.*, 1987; North *et al.*, 1990b). Using two-dimensional gels, 23 distinct activities have been identified (Neale & Alderete, 1990). Many of these enzymes, including those most active towards gelatin, have apparent molecular masses significantly larger than those of well-characterized CPs of higher eukaryotes. During axenic growth *in vitro*, *T. vaginalis* releases large quantities of proteinases (apparently exclusively of the cysteine type) into the growth medium (Lockwood *et al.*, 1987; North *et al.*, 1990b; Bóznér & Demeš, 1991). This has also been shown to occur *in vivo*, *T. vaginalis* CPs having been detected in vaginal washouts from infected women (Alderete *et al.*, 1991; Bóznér *et al.*, 1992). The extent to which these different trichomonad CPs represent separate gene products, rather than aggregates or modified forms of relatively few gene products, is unclear. However, differences in substrate specificity and inhibitor sensitivity

suggest that distinct enzymes are present (North *et al.*, 1990b) and that at least some of the multiplicity in enzyme activities could be due to multiple genes rather than post-translational modification or artifacts associated with the electrophoretic methods used for analysis. Until now, however, this had not been investigated at the molecular level.

Although some purification of trichomonad CPs has been achieved (Lockwood *et al.*, 1985, 1986; Garber & Lemchuk-Favel, 1989; Irvine *et al.*, 1993), it is only very recently that any primary sequence data have become available (Irvine *et al.*, 1993). Purification of trichomonad CPs has been problematic, and a molecular biology approach provides an alternative means of gaining information on the variation between the different CPs expressed by *T. vaginalis*. This paper describes the successful cloning of a number of cDNAs encoding CPs and the detailed characterization of four of them.

METHODS

Growth of parasites and preparation of DNA and RNA. A clonal cell line (G3) of *T. vaginalis* was grown in modified Diamond's medium (Diamond, 1957) as previously described (Lockwood *et al.*, 1984). The cells were harvested in the exponential phase of growth by centrifugation. For DNA preparation, the cells were washed twice with ice-cold TBS (25 mM Tris, 136 mM NaCl, 2.7 mM KCl) pH 7.4 (Sambrook *et al.*, 1989). DNA was extracted from the pellet using the method of Bowtell (1987). The DNA was dissolved overnight in 10 mM Tris/HCl, 10 mM EDTA, pH 8.0, then RNase A was added (to 100 µg ml⁻¹) and the mixture incubated at 37 °C for 1 h. SDS (to 0.5%) and proteinase K (to 100 µg ml⁻¹) were added and the samples incubated at 50 °C for 2 h followed by extraction with phenol/chloroform and then chloroform. This routinely gave DNA of 23.5 kb and beyond in size (based on gel electrophoresis) with 260/280 nm ratios of between 1.8 and 2.0. For RNA, the cells were washed once in ice-cold TBS and RNA was extracted from the pellet by a modification (Sambrook *et al.*, 1989) of the method of Chirgwin *et al.* (1979). Poly(A)⁺ RNA was isolated using oligo(dT) by the column method (Sambrook *et al.*, 1989) or using biotinylated oligo(dT) and streptavidin paramagnetic particles (PolyAtract mRNA Isolation System, Promega).

PCR and RACE-PCR. The degenerate oligonucleotide primers used to amplify the CP gene fragments were obtained from Dr J. C. Mottram (Glasgow) and were based on consensus sequences of highly conserved areas flanking the active-site cysteine and asparagine (North *et al.*, 1990a). The primers were synthesized with inosines in positions where all four bases were possible in a codon and with *Eco*RI recognition sites added to the 5' end of each primer to facilitate cloning. The sequences

were as follows (the IUB Group Codes for the redundancies are used and the cloning sites are underlined throughout):

primer 1,
5' CCGAATTCCARGGICARTGYGGIWSITGYTGG 3';
primer 2,
5' CCGAATTCCAISWRTTYTTIACDATCCARTA 3'.

These primers were used in PCR reactions with *T. vaginalis* genomic DNA as the target DNA in 100 µl reactions using the Perkin-Elmer Cetus GeneAmp kit components. Each reaction contained 100 ng trichomonad DNA and each of the primers at a final concentration of 0.5 µM. The reaction was carried out for 30 cycles with conditions as follows: 94 °C, denaturing, 1 min; 55 °C, annealing, 1 min; 72 °C, extension, 3 min. Appropriate controls were included in which PCR reactions contained the primers alone, either singly or together, without the target DNA. The PCR products were analysed by electrophoresis on 4% (w/v) agarose gels (3:1, wide range/standard, Sigma).

For RACE-PCR (Loh *et al.*, 1989), total cellular RNA from *T. vaginalis* was transcribed into single-stranded cDNA (ss cDNA) using the components of the 5' RACE System Kit (Life Technologies, according to the manufacturer's instructions) and a 21-mer primer complementary to all of the 3' untranslated region of the RNA corresponding to the cDNA clone TvCP1 (see below) apart from the stop codon. The primer sequence was as follows:

primer 3, 5' TGAACAGTAATATTTTATAGGT 3'.

The RACE-PCR procedure was performed with a 5' RACE System Kit according to the manufacturer's instructions based on the method of Loh *et al.* (1989). Briefly, excess dNTPs and primer were removed from the ss cDNA and a homopolymeric tail of dCs added to the 3' end of the ss cDNAs using terminal deoxynucleotidyl transferase. The ss cDNA was heated at 94 °C for 5 min, quenched on ice and 5 µl aliquots added to 100 µl PCR reactions containing Perkin-Elmer Cetus Kit components and an anchor primer (supplied with the kit) which included a sequence complementary to the poly dC tail:

primer 4,
5' CUACUACUACUAGGCCACGCGTCGACTAGTACG
GIIGGGIIGGGIIG 3'

and either primer 2 (see above) or primer 5, which was complementary to the entire 3' untranslated region of the cDNA clone TvCP1:

primer 5,
5' ACCGAATTCTGAACAGTAATATTTTATAGGTTTA 3'

each at a final concentration of 1.0 µM. The reaction conditions were as follows: 94 °C, denaturing, 1 min; 40 °C, annealing, 1 min; 72 °C, extension, 3 min (for 5 cycles); then as above except with an annealing temperature of 55 °C (for a total of 25 cycles). Appropriate controls were included as above, and in addition, PCR reactions were run with samples prepared without reverse transcriptase (no cDNA synthesis) or without terminal deoxynucleotidyl transferase (no tailing). The products of these reactions were run on 1% (w/v) agarose gels and bands of interest excised from the gel and directionally cloned into pSport 1 (Life Technologies).

Screening of the *T. vaginalis* λZap II cDNA library. A *T. vaginalis* cDNA library was constructed from total cellular RNA in λZap II (Stratagene) according to the manufacturer's instructions. A cloned genomic fragment p43 obtained by PCR (see above) was radiolabelled with [α -³²P]dCTP by random hexamer priming (Life Technologies) and used to screen the

library at moderate stringency. This entailed four 5 min washes at room temperature in 2 × SSC (1 × SSC is 150 mM NaCl, 15 mM sodium citrate), 0.1% SDS, and a final stringency wash in 2 × SSC, 0.1% SDS at 60 °C for 60 min. Approximately 50 hybridizing plaques were picked in the first screen and the ten most strongly hybridizing plaques were re-screened. This was repeated through another screen and the ten most strongly hybridizing, well-isolated, single plaques were picked. The phagemids were rescued and examined further by restriction mapping and partial sequencing. In addition, positives from the first primary screen of the λZap II cDNA library (53 in total and excluding those already worked up) were pooled and rescreened using the cDNA clone p101 isolated in the first screen, which corresponds to TvCP1 (see Results). This second screen was performed at higher stringency than the first and entailed a final stringency wash in 1 × SSC, 0.1% SDS at 68 °C for 60 min.

DNA sequencing. All sequencing was double-stranded and was carried out using the Sanger method (Sanger, 1977) with Sequenase enzyme (United States Biochemical). Partial sequencing of the 5' and 3' ends of inserts in pUC18 was performed using pUC forward and pUC reverse primers, respectively, whereas for inserts in pSport 1, pUC forward and T7 primers were used, and for those in pBluescript II, T3 and T7 primers were used. All primers were obtained from Promega. Clones of interest were sequenced completely in both directions using the above primers appropriate for the vector and custom oligonucleotide primers designed for their sequences. Overlapping sequences were combined to give complete sequences and restriction maps and the predicted amino acid sequences determined using the computer program DNASIS (Pharmacia).

Amino acid sequence alignment. Alignment of the amino acid sequences was performed using the MultAlin program (Cherwell Scientific), which is based on the algorithm described by Lipman & Pearson (1985), with small adjustments being made by eye. Comparisons of the amino acid sequences with other known CP sequences were carried out using the Fasta program available on the University of Wisconsin GCG package and by searching the Owl composite protein data base which was accessed through the SERC facility at Daresbury (UK).

Southern and Northern blotting. *T. vaginalis* genomic DNA was digested (10 µg DNA per digest) with the restriction endonucleases *Eco*RI, *Xho*I and *Xba*I (Life Technologies), none of which cut within the cDNAs analysed. The products of the reactions were run on a 0.8% agarose gel (Life Technologies) and blotted onto Hybond-C extra nitrocellulose (Amersham) as described in Sambrook *et al.* (1989). *T. vaginalis* poly(A)⁺ RNA (2.5 µg per well) was denatured and run on 1.4% (w/v) agarose/formaldehyde gels (Davis *et al.*, 1986) and blotted onto Hybond-C extra nitrocellulose as described by Sambrook *et al.* (1989). All blots were prehybridized for 2–3 h at 37 °C in 50% (v/v) formamide, 6 × SSC, 5 × Denhardt's solution, 0.5% SDS, sonicated and denatured salmon sperm DNA (100 µg ml⁻¹) and poly(A)⁺ RNA (1 µg ml⁻¹, Boehringer) and then hybridized for 20–24 h in the same mix with the entire insert of one of TvCP1, TvCP2, TvCP3 or TvCP4, labelled with [α -³²P]dCTP by random hexamer priming (Life Technologies). Following hybridization the blots were washed twice at room temperature, once in 2 × SSC, 0.5% SDS for 5 min and once in 2 × SSC, 0.1% SDS for 15 min, then in 0.1 × SSC, 0.1% SDS at 37 °C for 30 min, with a final stringent wash in 0.1 × SSC, 0.1% SDS at 68 °C for 30 min. The blots were exposed to X-ray film (X-OMAT AR, Kodak) with intensifying screens (CAWO Special, CAWO).

RESULTS

PCR amplification of genomic DNA encoding CPs

The starting point for the isolation of cDNAs encoding *T. vaginalis* CPs was the amplification by PCR of genomic DNA fragments using primers based on two well-conserved sequences found in most eukaryotic CPs. This approach has been used successfully with a number of other species of parasitic protozoa (Baylis *et al.*, 1992; Eakin *et al.*, 1992; Rosenthal & Nelson, 1992; Souza *et al.*, 1992; Traub-Cseko *et al.*, 1993). A *T. vaginalis* DNA fragment of approximately 500 bp was specifically and exclusively amplified and was present only when both primers were included in the PCR reaction (not shown). This was excised from the gel, cloned into pUC18 and a number of recombinant clones selected for analysis. On the basis of restriction mapping, 23 clones were placed into five groups. Examples of each were partially sequenced to establish their identity. At least three different sequences encoding proteins with significant similarity to CPs were present. One sequence was well represented amongst the clones and a 520 bp cloned PCR fragment from one of these clones, p43, was sequenced (sequence not shown) and used as a probe to screen a *T. vaginalis* cDNA library at moderate stringency.

Isolation of CP cDNA clones

The *T. vaginalis* λ Zap II cDNA library was screened under moderate stringency conditions using p43. Ten strongly hybridizing plaques were chosen from the final screen and the phagemids recovered. The clones were initially analysed by restriction mapping and then the 5' and 3' ends of the cloned DNA sequenced. The predicted amino acid sequences showed that they all encoded putative CPs and a total of five different CPs were represented. Interestingly, none of the CP sequences was identical to the original genomic PCR fragment. Based on comparisons with the amino acid sequences of known CPs, two of the clones, p101 and p102, appeared to be full-length. Five of the ten clones were identical to p101 and were either the same length or shorter. Clone p102 was the only clone with its sequence among the ten clones. Another clone, p103, appeared to be almost full-length (it was of similar length to the corresponding mRNA; see Fig. 4) and it too was the only clone with its sequence. The three remaining clones (p104, p105 and p106) encoded two other CPs that were very similar to each other, one represented by p104, the other by p105 and p106. The inserts for these three were either considerably less than a full-length cDNA or appeared to consist of two identical head-to-tail cDNA inserts (data not shown) and none of them was characterized further.

One of the clones, p201, isolated from a second library screen was identical to p101 and was chosen to represent this cDNA insert and designated TvCP1. Clone p102 was chosen to represent the second cDNA insert and designated TvCP2. p103, the nearly full-length clone, was designated TvCP4.

Some unusual features in the predicted amino acid sequences of both clone p101 and clone p102 suggested that, despite other evidence to the contrary, they might not be full length (see below). Two approaches were therefore used to clarify the situation with respect to p101, the shorter of the two clones. In the first of these, positives from the primary screen of the *T. vaginalis* λ Zap II cDNA library were pooled and rescreened using p101 cDNA under more stringent conditions to see if it was possible to isolate a longer clone. Ten strongly hybridizing plaques were again chosen from the final screen and the phagemids rescued. Restriction mapping and sequencing of the clones indicated that six were identical to p101. None of these clones was longer than the original p101. One clone out of the remaining four identified with p102 and sequencing confirmed that it was also the same length as the original clone. Based on restriction mapping, the other clones did not identify with either p101 or p102. As they were also much shorter (600–800 bp), they were not characterized further. The second approach involved a RACE-PCR procedure which should have amplified the complete 5' end of any cDNA related to p101. For the primer combination 4 and 5, the largest band which was specifically amplified was approximately 1100 bp. Five clones were picked at random and their 5' and 3' ends examined. They all identified with p101 and their longest open reading frames corresponded exactly to the longest open reading frame of this clone. Based on the above evidence, we concluded that clones p101 and p102 were almost certainly full-length.

Using RACE-PCR with primer 4 (the anchor primer) and one of the internal degenerate primers (primer 2, which was used to amplify the original genomic PCR fragment), bands of approximately 800 bp and 1200 bp were specifically and exclusively amplified. Several clones from the 800 bp and 1200 bp bands were sequenced, at least partially, but with two exceptions their predicted amino acid sequences did not identify with a CP (data not shown). Of the two that did encode CPs, both of which were from the 800 bp band, only one was close to being full length; this represented another distinct sequence and was designated TvCP3.

Predicted amino acid sequences of TvCP1–4

Two of the *T. vaginalis* CP cDNA clones (TvCP1 and TvCP2) analysed in detail were full length. Although neither of the other two (TvCP3 and TvCP4) was full length, it was useful to include them in the sequence analysis. The four clones were sequenced completely in both directions. The sequencing strategy and restriction map for each is shown in Fig. 1. As the nucleotide sequences *per se* are not the subject of this paper they are not included here, although all are available in the EMBL database. The clones were 1021 bp (TvCP1), 1086 bp (TvCP2), 861 bp (TvCP3) and 953 bp (TvCP4) in length. The longest open reading frame for TvCP1 predicted a 309 amino acid, 34.4 kDa protein and that for TvCP2, a 314 amino acid, 34.6 kDa protein, assuming in each case that the first methionine residue represents the N-terminal

Fig. 1. Restriction maps and sequencing strategy for TvCP1-4 cDNA inserts. Arrows indicate the length and direction of sequencing. The sequences of the protein coding regions are boxed. Letters represent restriction sites: C, *Clal*; D, *Dral*; Ha, *HaeIII*; Hd, *HindIII*; Sl, *Sall*; Sp, *Sspl*; St, *Styl*.

Table 1. Identity between *T. vaginalis* CPs

Identity was calculated as the percentage of the amino acid residues in each CP (top) for which an identical amino acid is present in the other CPs (left). Sections of the sequence equivalent to those not yet available for TvCP3 (the C-terminus) and TvCP4 (the N-terminus) were not included when comparisons were made with the latter sequences. In all other cases, all the amino acids in a sequence were included.

CP	Identity (%)			
	TvCP1	TvCP2	TvCP3	TvCP4
TvCP1	—	60	53	73
TvCP2	61	—	59	64
TvCP3	55	60	—	57
TvCP4	72	65	56	—

Table 2. Identity between TvCP1 and TvCP2 and CPs of other organisms

Identity was calculated as the percentage of the total amino acid residues in TvCP1 or TvCP2 for which an identical amino acid was present in the other CPs.

Proteinase	Identity (%)		Reference*
	TvCP1	TvCP2	
<i>Dictyostelium discoideum</i> CP1	37	33	1
<i>Dictyostelium discoideum</i> CP2	45	43	2
<i>Trypanosoma brucei</i> CP	36	33	3
Cruzipain (<i>T. cruzi</i>)	35	36	4
<i>Leishmania mexicana</i> CPa	36	33	5
<i>Leishmania mexicana</i> CPb	34	32	6
Cathepsin L (human)	42	37	7
Cathepsin S (human)	39	35	8
Cathepsin H (human)	37	35	9
Lobster LCP1	40	39	10
Lobster LCP2	43	40	10
Lobster LCP3	41	40	10
Oilseed rape COT 4	40	40	11
Black gram SH-EP	40	40	12
Pea (clone 15a)	36	35	13
Oryzain α (rice)	41	42	14
Oryzain β (rice)	41	39	14

*1, Williams *et al.* (1985); 2, Pears *et al.* (1985); 3, Mottram *et al.* (1989); 4, Campetella *et al.* (1992); 5, Mottram *et al.* (1992); 6, Souza *et al.* (1992); 7, Gal & Gottesman (1988); 8, Shi *et al.* (1992); 9, Fuchs & Gassen (1989); 10, Laycock *et al.* (1991); 11, Dietrich *et al.* (1989); 12, Akasofu *et al.* (1989); 13, Guerrero *et al.* (1990); 14, Watanabe *et al.* (1991).

amino acid (see below). TvCP3 and TvCP4 have open reading frames of 278 and 292 amino acids, respectively. All four are more closely related to one another (Table 1) than they are to CPs from other organisms (see Table 2 for TvCP1 and TvCP2). TvCP1 and TvCP2 both show greatest similarity to TvCP4, whereas TvCP3, the most divergent of the four, is most similar to TvCP2. Identity searches showed that the proteins encoded by the four clones belonged to the cathepsin L/cathepsin H/papain branch of the papain superfamily. *Dictyostelium discoideum* CP2 is the closest of the non-trichomonad CPs to TvCP1, TvCP2, TvCP4 (45% identical residues in CP2) and TvCP3 (38% identity). TvCP1, TvCP2 and TvCP4 were also relatively close to mammalian cathepsin L, three lobster CPs and CPs from rice (oryzains α and β), oilseed rape and black gram. In every case the degree of identity was considerably greater in that part of the sequence corresponding to the central region (45–56% identity) than in that corresponding to the prepro-region (15–28%). The trichomonad CPs were less similar to CPs of other protozoa, including those of other flagellates (*Trypanosoma brucei*, *T. cruzi* and *Leishmania mexicana*) (32–36% identity for TvCP1, TvCP2 and TvCP4; 26–31% for TvCP3). All four trichomonad sequences

TvCP1	MMYQAHEQKSFGLGWMRETGNMFTGDEYHQRFG	*
TvCP2	MF-AFLSGATSNVLKHEEKAFLAYMRETGNFTGDEYHFRIG	
TvCP3	MFAFFATASSKFLQHEEKAFLDWMRSTNNMFGDEYHFRIG	
TvCP4	WMRETGNMFTGDEYHQRFG	
HCL	MNPTLILAAFLGCLGIASATLTFDHSLEAQWTKWKAMHNRLYGMNEEGWRRRA	
	-113 -100	
TvCP1	IWLSNKRLLVQHNAA----NGGFVLAMNKLHLSPSEYKALLGFKNEKRS	*
TvCP2	IYLANKRLLVQHNAA----NKGFKLGLNKLHLTQSEYRSLGAKRLG-Q	*
TvCP3	VYNTNKRRLVQHNRA----NSGYQLTMNHLSCMTPSEYKVLGKHQTKK-	*
TvCP4	IWLSNKRLLVQHNRA----NLGFTVALNKLHLTPAEYNSLLGFRMKN-A	*
HCL	VWEKNMKMIELHQEYREGKHSFTMAMNAFGDMTSEEFQVMNGFQNRKP	
	-50	
TvCP1	D-RVKPIASN--YVAPASIDWREKGVVNEIKDQGGCGSCWTFSTIQAMES	***
TvCP2	K-SGNFFKCD--APANDAVDWRDKGIVNFKDQGGCGSCWAFSAIQASES	*
TvCP3	EGEAKIFKGD---VPAVDWRNAKIVNFKDQGGCGSCWAFSAIQASES	*
TvCP4	E-R-KAVKSN--AIANACDWRKKGAVNFKDQGGCGSCWAFSAIQASES	*
HCL	R-KGKVFQEPFLFYEAAPRSVDWREKGVVTFVKNQGGCGSCWAFSATGALEG	*
	1 Δ	
TvCP1	QWAVKHTKLYSLSEQLNLDVCTTC--YGCNGLMELAYDYKYQKGF	***
TvCP2	RYAQANKQLLDLAELQIVDCTTC--YGCNGLMELAYDYKYQKGF	*
TvCP3	QWALKKQLLQSLAEQNMVDCVDTTC--YGCNGLMELAYDYKYQKGF	*
TvCP4	QYYSFKTLQSLSEQLNLDVCTTC--YGCNGLMELAYDYKYQKGF	*
HCL	QMRKTKRLISLSEQLNLDVCTTC--YGCNGLMELAYDYKYQKGF	*
	50	
TvCP1	TEADYPYKAIDQSCCKFNAKVAEPTVTGYITVTE-GDEKDLMNKVAQYGP	*
TvCP2	LTADYPYTARDGTCKFHASK-SVGLTKGYDEVKD--TEAEL-AKAASKGV	*
TvCP3	LETADYPYTARDGSCCKFNAK-GVTLTKSYVRPTTQNEDELKAGCAKGV	*
TvCP4	TEADYPYTARDGSCCKFNAK-GTSQIKSYVNVAE-GDEKDLATKVSITLGP	*
HCL	SEESYPYEATEESCKYNP-KYSVANDTGFVDIPK--QEKALMKAVATVGP	*
	100	
TvCP1	AAIAIDASHYSFQLYSSGIYDESSCSPEGLDHAVGCVGSGSEG----	***
TvCP2	VSVCIDASHYSFQLYTSGIYDESSCSAWNLDHAGLVGYGTGEG----	*
TvCP3	VSIADASHYSFQLYSSGIYDESSCSSTFLDHAGLVGYGTGEG----	*
TvCP4	AAIAIDASHYSFQLYSSGIYDESSCSSTFLDHAGLVGYGTGEG----	*
HCL	ISVAIDAGHSFQLYSSGIYDESSCSSTFLDHAGLVGYGTGEG----	*
	150 Δ	
TvCP1	YWIVRNSWGSWGEGKYIRMIKDKNNQCGEASAACTPTVSA	*****
TvCP2	YWIVRNSWGSWGEGKYIRMIKDKNNQCGEASAACTPTVSA	*
TvCP4	YWIVRNSWGSWGEGKYIRMIKDKNNQCGEATMACIPQDK	*
HCL	YWLKNSWGEWGMGGYVMAKDRNRHCGIASAASYPTV	*
	Δ 200	

Fig. 2. Comparison of the predicted amino acid sequences of *T. vaginalis* cysteine proteinases with each other and with human cathepsin L (HCL; Gal & Gottesman, 1988). Δ below a residue indicates key active-site residues. * above a residue indicates a match across all sequences. Dashes indicate gaps made to maximize alignments. The numbers refer to the numbering system of HCL. The nucleotide sequences of cDNAs TvCP1, TvCP2, TvCP3 and TvCP4 have been submitted to EMBL and have been assigned the accession numbers X77218, X77219, X77220 and X77221, respectively.

showed much less similarity to mammalian cathepsin B (19–24% identity) and thus resembled most other reported protozoan CPs (North, 1992). Only one protozoan CP whose sequence has cathepsin B-like features has been described (Robertson & Coombs, 1993).

Fig. 2 shows the four *Trichomonas vaginalis* sequences aligned with each other and cathepsin L (Gal & Gottesman, 1988). The predicted products of the trichomonad cDNAs had most of the features typical of eukaryotic CPs. Within the central region (North *et al.*, 1990a) corresponding to the mature form of mammalian cathepsins and plant CPs [residues 1–220, human cathepsin L (HCL) numbering is used throughout] every residue known to be essential for catalytic activity is present in all

four sequences. They all have the active-site residues cysteine-25, histidine-163 and asparagine-187 (for TvCP3 this latter residue is not shown, as the corresponding sequence covers the area to which the degenerate PCR primers hybridize). They also had the six conserved cysteine residues which form three disulphide bonds in other CPs (HCL residues 22, 56, 65, 98, 156, 209) and the majority of the other residues which are highly conserved amongst CPs. There are some interesting distinctions, however, in the positions corresponding to those of the S_2 substrate-binding site in papain (Baker & Drenth, 1987). Residue 70, which is a methionine in cathepsin L, is also a methionine in TvCP1 and TvCP4, but is proline in TvCP2 and glutamic acid in TvCP3. Cathepsin L has an alanine residue at position 214 as does TvCP1, while TvCP4 has a methionine. In contrast, TvCP2 has a glutamic acid residue. Differences between the other residues comprising the S_2 subsite are less marked.

There are some other noteworthy features within the central region. First, in neither TvCP2 nor TvCP4 is there a proline residue at the position corresponding to the second residue in most mature CPs (including cathepsin L, Fig. 2), although TvCP2 has a proline at the –2 position. Proline is found almost universally at position 2, but exceptions have been found recently in lobster CPs (Laycock *et al.*, 1991), some CPs from *D. discoideum* (A. Champion, A. Gooley, M. J. North & K. L. Williams, unpublished) and other trichomonad CPs (Irvine *et al.*, 1993). Second, in the region corresponding to cathepsin L residues 77–78 there are two additional residues in each of the sequences. This is also true for other *T. vaginalis* CP sequences, namely the genomic PCR fragment clone p43 and the cDNA clone p106 (data not shown), and for *Trichomonas foetus* CP sequences (D. J. Mallinson & M. J. North, unpublished). While small differences in the number of residues in particular sections of CP sequences are by no means exceptional, it is interesting to note that there are two extra residues in this position in the sequences of all other flagellate CPs that have been sequenced (Mottram *et al.*, 1989, 1992; Campetella *et al.*, 1992; Souza *et al.*, 1992; Traub-Cseko *et al.*, 1993) but not in any other reported CP sequences. Third, trichomonad sequences have a cysteine residue at position 60. This is present in all the *Trichomonas vaginalis* clones, including clones p43, p104 and p106 (data not shown), and also in most *Trichomonas foetus* CPs (D. J. Mallinson & M. J. North, unpublished). A cysteine is found in this position in a pea CP (Guerrero *et al.*, 1990) and *D. discoideum* CP1 (Williams *et al.*, 1985) but in no others reported to date.

All four *Trichomonas vaginalis* CP genes encode a typical pro-region (cathepsin L residues –95 to –1) which in other CPs is removed during activation of the pro-enzyme. As among other groups of CP (Mottram *et al.*, 1989; Souza *et al.*, 1992), the pro-regions are more diverged than the central regions but all show some similarity to other CP pro-regions. Unlike some of the trypanosomatid CPs (Mottram *et al.*, 1989; Campetella *et al.*, 1992; Souza *et al.*, 1992), the *T. vaginalis* CPs do not

have a C-terminal extension. For TvCP1, TvCP2 and TvCP4 this conclusion is based on the predicted amino acid sequences. The TvCP3 cDNA clone does not include the region corresponding to the C-terminus of the enzyme, but the length of the mRNA (see Fig. 4) is inconsistent with an extended coding region at the 3' end.

The most striking feature of the TvCP1, TvCP2 and TvCP3 sequences (this part of the sequence was not available for TvCP4) is at the N-terminus (Fig. 2). All the available evidence suggests that trichomonad CPs are lysosomal (Lockwood *et al.*, 1988) and as most secretory and lysosomal proteins have a pre-sequence corresponding to a signal peptide it was expected that the trichomonad sequences would each have encoded one. A signal sequence characteristically has one or more positively-charged amino acids near its N-terminus, followed by a continuous stretch of 10 ± 3 hydrophobic residues (Gierasch, 1989). A signal peptidase cleavage site, with small residues at positions -1 and -3 (with respect to the cleaved bond), is normally found 4-6 residues from the C-terminal end of the hydrophobic region (Dalbey & von Heijne, 1992). TvCP1 lacks any such sequence. RACE-PCR independently confirmed that the open reading frame does not extend beyond the methionine shown in Fig. 2. TvCP2 and TvCP3 appear to have very short pre-sequences in which, after the N-terminal methionine, there is a short stretch (5 residues) of hydrophobic (TvCP2) or mostly hydrophobic (TvCP3) residues followed immediately by three residues (S-G-A at positions -99 to -97 in TvCP2, A-T-A at positions -100 to -98 in TvCP3) which define possible cleavage sites. In TvCP2 the open reading frame could not extend beyond the methionine indicated since the codon for the latter was preceded by an in-frame termination codon (UAA). The first methionine codon in the TvCP3 sequence, which was obtained from the longest RACE-PCR product, aligned with that of TvCP2.

The two most closely related clones, TvCP1 and TvCP4, share some features not apparent in the other two clones. In particular they both have cysteines at positions 167 and 212. This pair is also apparent in clones p104 and p106 (data not shown).

Many other CPs, including those of other flagellates, are glycosylated or have potential *N*-glycosylation sites in the pro-region and/or central region. No such sites are apparent in any of the *T. vaginalis* clones.

Genomic organization and expression of TvCP1-4

Trichomonas vaginalis genomic DNA was digested with three restriction enzymes that do not cut within either TvCP1 and TvCP2 or within the sequences known for TvCP3 and TvCP4 and probed in Southern blots with each of the CP cDNAs (Fig. 3). With the exception of TvCP4, all the clones hybridized to single bands in every case, which suggests that the corresponding genes are present in the genome as single copies. TvCP4 hybridized strongly to three bands in *EcoRI* and *XbaI* digests,

Fig. 3. Southern blot analysis of TvCP1-4 genes. *T. vaginalis* genomic DNA (10 µg) was digested with restriction enzymes, electrophoresed on a 0.8% agarose gel, transferred to nitrocellulose and hybridized individually with either TvCP1 (lanes 1-3), TvCP2 (4-6), TvCP3 (7-9) or TvCP4 (10-12). Digestion with *EcoRI* (1, 4, 7, 10), *XbaI* (2, 5, 8, 11), *XhoI* (3, 6, 9, 12). Marker sizes (Life Technologies) are shown on the right.

suggesting that the corresponding gene is present as multiple copies. The possibility that this may represent cross-hybridization with other very similar, but as yet uncharacterized, genes cannot be ruled out at present. For example, the nucleotide sequences available for the uncharacterized clones p104 and p106 (see above) have over 88% identity with TvCP4, but the clones must represent the transcripts of different genes as their 3' untranslated regions were distinct (data not shown).

Fig. 4. Northern blot analysis of TvCP1-4 gene expression. Poly(A)⁺ RNA (2.5 µg) was electrophoresed on a 1.4% (w/v) agarose/formaldehyde gel, transferred to nitrocellulose, and hybridized individually with either TvCP1 (lane 1), TvCP2 (lane 2), TvCP3 (lane 3) or TvCP4 (lane 4). RNA ladder sizes (Life Technologies) are shown on the right.

There were found to be differences in the organization of the four genes corresponding to TvCP1–4. TvCP1 and TvCP2 gave quite similar but reproducibly different restriction patterns. The restriction patterns for TvCP3 and TvCP4 were very different from one another and from those of TvCP1 and TvCP2. Despite being similar to one another in sequence, the four cDNAs did not cross-hybridize under the conditions employed.

Northern blots of poly(A)⁺ RNA were probed with each of the CP cDNAs (Fig. 4). All four genes were expressed at similar levels as abundant, single transcripts which were also easily detectable in total RNA (data not shown). The transcripts were detectable as similar-sized, diffuse bands. The lengths of the mRNAs for TvCP1 and TvCP2 (1000–1200 bp) both corresponded well to the length of their respective cDNAs, indicating that they are probably full-length. For TvCP4 (960–1120 bp) a comparison of the length of the mRNA with the cDNA insert indicated that it is nearly full-length. The TvCP3 message is of similar size (1000–1150 bp) to the other mRNAs.

DISCUSSION

In this report we have described the isolation and identification of seven DNA molecules which encode different CPs in *T. vaginalis*. Six were cDNAs and so the corresponding genes must have been expressed by the parasite. For the seventh, obtained by PCR amplification of genomic DNA, no corresponding longer or full-length cDNA was isolated from the cDNA library and so this gene may not be expressed under the growth conditions used. On the basis of these results, therefore, it is possible to state that at least six expressed CP genes are present in *T. vaginalis*. This provides an explanation for at least some of the multiplicity in CP activities in the parasite (Lockwood *et al.*, 1987; Neale & Alderete, 1990; North *et al.*, 1990b). Post-translational modification of one or more gene products may create further multiplicity, although the sequence data reported here suggest that differences created by N-linked glycosylation are not involved as no putative N-glycosylation sites are present. The possibility that there are more CP genes cannot be ruled out, and the presence of additional CP genes would help to account for the 23 proteinases reported to date (Neale & Alderete, 1990). Very recently, the N-terminal amino acid sequence of a purified *T. vaginalis* CP (23 kDa) has become available (Irvine *et al.*, 1993). This does not correspond to any of TvCP1, TvCP2, TvCP3 or TvCP4 and so must be a product of a different gene. It could be encoded by the gene represented by the genomic PCR fragment or either of the cDNAs which have not been characterized (the sequences over the corresponding region are not yet available). However, it does further emphasize the great multiplicity of CP genes in trichomonads.

The sizes predicted for the enzymes in this study are similar to those of the well-characterized CPs of other organisms. Assuming that the putative pro-region is cleaved, all would have molecular masses of approximately 24 kDa. This is less than the apparent molecular

masses of most of the *T. vaginalis* enzymes detected on electrophoresis gels (Lockwood *et al.*, 1987; North *et al.*, 1990b), although a recently purified CP did have a molecular mass of approximately 23 kDa (Irvine *et al.*, 1993). The reasons why the majority of the *T. vaginalis* CPs reported so far have unusually high apparent molecular masses, as determined from their electrophoretic mobility in partially denaturing gels, are not revealed by the sequence data presented here.

The *T. vaginalis* CPs detected in parasite lysates display differing substrate preferences and three types of specificity have been observed (North *et al.*, 1990b). It is of interest to consider the predicted make up of the S₂ binding site of the cloned *T. vaginalis* CPs. For TvCP1 and TvCP4, this site is hydrophobic in character and very closely resembles that of mammalian cathepsin L. Thus they would be expected to hydrolyse substrates with hydrophobic residues at the P₂ position, and two of the specificity groups defined earlier (North *et al.*, 1990b) do this. The other two CPs, TvCP2 and TvCP3, had a glutamic acid residue in the putative S₂ binding site. An acidic residue in the position corresponding to cathepsin L residue 214, as found in TvCP2, has been linked to an ability to hydrolyse substrates with arginine at the P₂ position, a feature of mammalian cathepsin B (Hasnain *et al.*, 1993), cruzipain (the major CP of *Trypanosoma cruzi*; Campetella *et al.*, 1992) and *Entamoeba* CPs (Scholze, 1991). *Trichomonas vaginalis* CPs with this specificity have been detected (North *et al.*, 1990b). Future work will be directed at identifying the products of the TvCP genes to confirm these links between structure and specificity.

The most intriguing feature of the CP sequences is the apparent lack of typical pre-sequences expected for lysosomal enzymes. Indeed TvCP1 has no recognizable pre-sequence, while in TvCP2 and TvCP3 the putative pre-sequences are very short. The significance of this remains to be elucidated. It would seem unlikely that the mechanism of synthesis of lysosomal proteins of trichomonads differs greatly from that in other eukaryotic cells (Kornfeld & Mellman, 1989), although it is possible that unique signals might be involved for targeting proteins to the lumen of the endoplasmic reticulum in these parasites. The short hydrophobic regions of TvCP2 and TvCP3 which precede conventional signal peptidase cleavage sites may be enough for targeting these proteins. TvCP1 may be targeted in a similar way to ovalbumin. This also lacks a cleavage site (Leader, 1979), although it does have a moderately hydrophobic N-terminus which is not apparent in TvCP1. At present there is a paucity of information on the mechanism of protein biosynthesis and targeting in trichomonads. The only other proteins of *T. vaginalis* whose targeting has been studied are ferredoxin and β -succinyl-coenzyme A synthetase. Both are nuclear-encoded and transported into hydrogenosomes, membrane-bound organelles involved in energy metabolism (Müller, 1993), and both have unusually short N-terminal pre-sequences when compared with those known to direct proteins to mitochondria and chloroplasts in other organisms (Johnson *et al.*, 1990; Lahti *et al.*, 1992).

Unlike lysosomal enzymes, however, these proteins are synthesized on free ribosomes, so they are not directly comparable.

On the basis of rRNA sequencing, it has been proposed that *T. vaginalis* and other trichomonads represent an early branch of the eukaryotic tree (Baroin *et al.*, 1988; Sogin, 1989). Three of the cloned trichomonad CPs (1, 2 and 4) are similar to a number of CPs from diverse origins (plants, vertebrates, invertebrates). This suggests that an ancestral CP appeared very early in the evolution of eukaryotic organisms. The high degree of relatedness of the four *T. vaginalis* CPs, however, suggests that the corresponding genes arose as a result of gene duplication which has occurred since the divergence of trichomonads from other eukaryotes. Three of the TvCP genes are single copy and only the TvCP4 gene may be multiple-copy. Some parasitic protozoa have multiple-copy CP genes arranged in tandem array (Mottram *et al.*, 1989; Campetella *et al.*, 1992; Souza *et al.*, 1992; Traub-Cseko *et al.*, 1993), with the individual genes being identical or very closely related. These genes encode products with a C-terminal extension, which is lacking from TvCP4, and the large number of copies correlates with relatively high levels of expression. This is not the case for TvCP4, as all four TvCP genes were found to be expressed to similar levels.

The high levels of CP activity in *T. vaginalis*, their multiplicity and the continuous release from cells suggest that proteolysis is likely to be important in the life of this organism (North, 1991). The results reported here have provided the first detailed information on CP structure and the basis of CP multiplicity. Such information will not only be important for our understanding of how the individual enzymes function but will also help in assessing the similarities and differences between the parasite and host enzymes and thus whether individual parasite CPs could be potential targets for new anti-trichomonad drugs based on CP inhibitors or substrates. The apparently unusual nature of the N-terminal region of the TvCP gene products has raised the possibility that some features of protein targeting in trichomonads might be distinct from those of the host and so make this process a potential target too. There have been very few molecular studies on trichomonads so far but the results hold promise that future studies on these relatively primitive eukaryotes will reveal additional peculiarities in the processes involved in gene expression and the synthesis and targeting of proteins.

ACKNOWLEDGEMENTS

We would like to thank Dr Michael Leaver and Dr Paul Hodgson for help with computer analyses of the DNA and protein sequences. We would also like to thank Dr Jeremy Mottram for help and advice in making the *T. vaginalis* cDNA library. The work was supported by a grant from the Wellcome Trust.

REFERENCES

- Alderete, J. F., Newton, E., Dennis, C. & Neale, K. A. (1991). The vagina of women infected with *Trichomonas vaginalis* has numerous proteinases and antibody to trichomonad proteinases. *Genitourin Med* **67**, 469–474.
- Akasofu, H., Yamauchi, D., Mitsuhashi, W. & Minamikawa, T. (1989). Nucleotide sequence of cDNA for sulfhydryl-endopeptidase (SH-EP) from cotyledons of germinating *Vigna mungo* seeds. *Nucleic Acids Res* **17**, 6733.
- Baker, E. N. & Drenth, J. (1987). The thiol proteases: structure and mechanism. In *Biological Macromolecules and Assemblies*, vol. 3, *Active Sites of Enzymes*, pp. 312–368. Edited by J. McPherson. New York: Wiley.
- Baroin, A., Perasso, R., Qu, L. H., Brugerolle, G., Bachellerie, J. P. & Adoutte, A. (1988). Partial phylogeny of the unicellular eukaryotes based on rapid sequencing of a portion of 28S ribosomal RNA. *Proc Natl Acad Sci USA* **85**, 3474–3478.
- Baylis, H. A., Megson, A., Mottram, J. C. & Hall, R. (1992). Characterisation of a gene for a cysteine protease from *Theileria annulata*. *Mol Biochem Parasitol* **54**, 105–108.
- Bowtell, D. D. L. (1987). Rapid isolation of eukaryotic DNA. *Anal Biochem* **162**, 463–465.
- Bóznér, P. & Demeš, P. (1991). Proteinases in *Trichomonas vaginalis* and *Tritrichomonas mobilensis* are not exclusively of cysteine type. *Parasitology* **102**, 113–115.
- Bóznér, P., Gombosová, A., Valent, M., Demeš, P. & Alderete, J. F. (1992). Proteinases of *Trichomonas vaginalis*: antibody response to patients with urogenital trichomoniasis. *Parasitology* **105**, 387–391.
- Campetella, O., Henriksson, J., Åslund, L., Frasch, A. C. C., Pettersson, U. & Cazzulo, J. J. (1992). The major cysteine proteinase (cruzipain) from *Trypanosoma cruzi* is encoded by multiple polymorphic tandemly organized genes located on different chromosomes. *Mol Biochem Parasitol* **50**, 225–234.
- Chirgwin, J. M., Prybyla, A. E., MacDonald, R. J. & Rutter, W. J. (1979). Isolation of biologically active ribonucleic acid from sources enriched in ribonuclease. *Biochemistry* **18**, 5294–5297.
- Dalbey, R. E. & von Heijne, G. (1992). Signal peptidases in prokaryotes and eukaryotes – a new protease family. *Trends Biochem Sci* **17**, 474–478.
- Davis, L. G., Dibner, M. D. & Battey, J. F. (1986). *Basic Methods in Molecular Biology*. New York: Elsevier.
- Diamond, L. (1957). The establishment of various trichomonads of animals and man in axenic culture. *J Parasitol* **43**, 488–490.
- Dietrich, R. A., Masylar, D. J., Heupel, R. C. & Harada, J. J. (1989). Spatial patterns of gene expression in *Brassica napus* seedlings: identification of a cortex-specific gene and localization of mRNAs encoding isocitrate lyase and a polypeptide homologous to proteinases. *Plant Cell* **1**, 73–80.
- Eakin, A. E., Mills, A. A., Harth, G., McKerrow, J. H. & Craik, C. S. (1992). The sequence, organization, and expression of the major cysteine protease (cruzain) from *Trypanosoma cruzi*. *J Biol Chem* **267**, 7411–7420.
- Fuchs, R. & Gassen, H. G. (1989). Nucleotide sequence of human preprocathepsin H, a lysosomal cysteine proteinase. *Nucleic Acids Res* **17**, 9471.
- Gal, S. & Gottesman, M. M. (1988). Isolation and sequence of a cDNA for human procathepsin L. *Biochem J* **253**, 303–306.

- Garber, G. & Lemchuk-Favel, L. T. (1989). Characterization and purification of extracellular proteases of *Trichomonas vaginalis*. *Can J Biochem* **35**, 903–909.
- Gierasch, L. M. (1989). Signal sequences. *Biochemistry* **28**, 923–930.
- Guerrero, F. D., Jones, J. T. & Mullet, J. E. (1990). Turgor-responsive gene transcription and RNA levels increase rapidly when pea shoots are wilted. Sequence and expression of three inducible genes. *Plant Mol Biol* **15**, 11–26.
- Hasnain, S., Hirana, T., Huber, C. P., Mason, P. & Mort, J. S. (1993). Characterization of cathepsin B specificity by site-directed mutagenesis. Importance of Glu(245) in the S2-P2 specificity for arginine and its role in transition state stabilization. *J Biol Chem* **268**, 235–240.
- Heine, P. & McGregor, J. A. (1993). *Trichomonas vaginalis*: a re-emerging pathogen. *Clin Obstet Gynecol* **36**, 137–144.
- Irvine, J. W., Coombs, G. H. & North, M. J. (1993). Purification of cysteine proteinases from trichomonads using bacitracin Sepharose. *FEMS Microbiol Lett* **110**, 113–120.
- Johnson, P. J. (1993). Metronidazole and drug resistance. *Parasitol Today* **9**, 183–186.
- Johnson, P. J., d'Oliveira, C. E., Gorrell, T. E. & Müller, M. (1990). Molecular analysis of the hydrogenosomal ferredoxin of the anaerobic protist *Trichomonas vaginalis*. *Proc Natl Acad Sci USA* **87**, 6097–6091.
- Kornfeld, S. & Mellman, I. (1989). The biogenesis of lysosomes. *Annu Rev Cell Biol* **5**, 483–525.
- Kreiger, J. N., Verdon, M., Siegal, N. & Holmes, K. K. (1993). Natural history of urogenital trichomoniasis in men. *J Urol* **149**, 1455–1458.
- Lahti, C. J., d'Oliveira, C. E. & Johnson, P. J. (1992). β -Succinyl-coenzyme A synthetase from *Trichomonas vaginalis* is a soluble hydrogenosomal protein with an amino-terminal sequence that resembles mitochondrial presequences. *J Bacteriol* **174**, 6822–6830.
- Laycock, M. V., MacKay, R. M., Di Fruscio, M. & Gallant, J. W. (1991). Molecular cloning of three cDNAs that encode cysteine proteinases in the digestive gland of the American lobster (*Homarus americanus*). *FEBS Lett* **292**, 115–120.
- Leader, D. P. (1979). Protein biosynthesis on membrane-bound ribosomes. *Trends Biochem Sci* **4**, 205–208.
- Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* **227**, 1435–1441.
- Lockwood, B. C., North, M. J. & Coombs, G. H. (1984). *Trichomonas vaginalis*, *Tritrichomonas foetus* and *Trichomitus batrachorum*: comparative proteolytic activity. *Exp Parasitol* **58**, 245–253.
- Lockwood, B. C., North, M. J. & Coombs, G. H. (1985). Purification and characterization of proteinases of the parasitic protozoan *Trichomonas vaginalis*. *Biochem Soc Trans* **13**, 336.
- Lockwood, B. C., North, M. J. & Coombs, G. H. (1986). Proteolysis in trichomonads. *Acta Univ Carol Biol* **30**, 313–318.
- Lockwood, B. C., North, M. J., Scott, K. I., Bremner, A. F. & Coombs, G. H. (1987). The use of a highly sensitive electrophoretic method to compare the proteinases of trichomonads. *Mol Biochem Parasitol* **24**, 89–95.
- Lockwood, B. C., North, M. J. & Coombs, G. H. (1988). The release of hydrolases from *Trichomonas vaginalis* and *Tritrichomonas foetus*. *Mol Biochem Parasitol* **30**, 135–142.
- Loh, E. Y., Elliot, J. F., Cwirla, S., Lanier, L. L. & Davis, M. M. (1989). Polymerase chain reaction with single-sided specificity: analysis of T cell receptor δ chain. *Science* **243**, 217–220.
- McKerrow, J. H., Sun, E., Rosenthal, P. J. & Bouvier, J. (1993). The proteases and pathogenicity of parasitic protozoa. *Annu Rev Microbiol* **47**, 821–853.
- Mottram, J. C., North, M. J., Barry, J. D. & Coombs, G. H. (1989). A cysteine proteinase cDNA from *Trypanosoma brucei* predicts an enzyme with an unusual C-terminal extension. *FEBS Lett* **258**, 211–215.
- Mottram, J. C., Robertson, C. D., Coombs, G. H. & Barry, J. D. (1992). A developmentally regulated cysteine proteinase gene of *Leishmania mexicana*. *Mol Microbiol* **6**, 1925–1932.
- Müller, M. (1993). The hydrogenosome. *J Gen Microbiol* **139**, 2879–2889.
- Neale, K. A. & Alderete, J. F. (1990). Analysis of the proteinases of representative *Trichomonas vaginalis* isolates. *Infect Immun* **58**, 157–162.
- North, M. J. (1991). Proteinases of trichomonads and *Giardia*. In *Biochemical Protozoology*, pp. 234–244. Edited by G. H. Coombs & M. J. North. London: Taylor & Francis.
- North, M. J. (1992). The characteristics of cysteine proteinases of parasitic protozoa. *Biol Chem Hoppe Seyler* **373**, 401–406.
- North, M. J., Mottram, J. C. & Coombs, G. H. (1990a). Cysteine proteinases of parasitic protozoa. *Parasitol Today* **6**, 270–274.
- North, M. J., Robertson, C. D. & Coombs, G. H. (1990b). The specificity of trichomonad cysteine proteinases analysed using fluorogenic substrates and specific inhibitors. *Mol Biochem Parasitol* **39**, 183–193.
- Pears, C. J., Mahbubani, H. M. & Williams, J. G. (1985). Characterization of two highly diverged but developmentally co-regulated cysteine proteinase genes in *Dictyostelium discoideum*. *Nucleic Acids Res* **13**, 8853–8866.
- Robertson, C. D. & Coombs, G. H. (1993). Cathepsin B-like cysteine proteases of *Leishmania mexicana*. *Mol Biochem Parasitol* **62**, 271–280.
- Rosenthal, P. J. & Nelson, R. G. (1992). Isolation and characterization of a cysteine proteinase gene of *Plasmodium falciparum*. *Mol Biochem Parasitol* **51**, 143–152.
- Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: a Laboratory Manual*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci USA* **74**, 5463–5467.
- Scholz, H. (1991). Amoebapain, the major proteinase of pathogenic *Entamoeba histolytica*. In *Biochemical Protozoology*, pp. 251–256. Edited by G. H. Coombs & M. J. North. London: Taylor & Francis.
- Shi, G. P., Munger, J. S., Meara, J. P., Rich, D. H. & Chapman, H. A. (1992). Molecular cloning and expression of human alveolar macrophage cathepsin S, an elastolytic cysteine protease. *J. Biol Chem* **267**, 7258–7262.
- Sogin, M. L. (1989). Evolution of eukaryotic microorganisms and their small subunit ribosomal RNAs. *Am Zool* **29**, 487–499.
- Souza, A. E., Waugh, S., Coombs, G. H. & Mottram, J. C. (1992). Characterization of a multicopy gene for a major stage-specific cysteine proteinase of *Leishmania mexicana*. *FEBS Lett* **311**, 124–127.
- Traub-Cseko, Y. M., Duboise, M., Boukai, L. K. & McMahon-Pratt, D. (1993). Identification of two distinct cysteine proteinase genes of *Leishmania pifanoi* axenic amastigotes using the polymerase chain reaction. *Mol Biochem Parasitol* **57**, 101–116.

Watanabe, H., Abe, K., Emori, Y., Hosoyama, H. & Arai, S. (1991). Molecular cloning and gibberellin-induced expression of multiple cysteine proteinases of rice seeds (oryzains). *J Biol Chem* **266**, 16897-16902.

Williams, J. G., North, M. J. & Mahbubani, H. (1985). A developmentally regulated cysteine proteinase in *Dictyostelium discoideum*. *EMBO J* **4**, 999-1006.

Received 28 February 1994; revised 23 June 1994; accepted 30 June 1994.