Cathepsin F of Teladorsagia circumcincta is a recently evolved cysteine protease

Sarah Sloan, Caitlin Jenvey, Callum Cairns and Michael Stear

AgriBio Centre for AgriBioscience, Department of Animal, Plant and Soil Sciences, School of Life Sciences, La Trobe University, Bundoora, Victoria, Australia.

Evolutionary Bioinformatics Volume 16: 1-12 © The Author(s) 2020 Article reuse guidelines: sagepub.com/journals-permissions DOI: 10.1177/1176934320962521



ABSTRACT: Parasitic cysteine proteases are involved in parasite stage transition, invasion of host tissues, nutrient uptake, and immune evasion. The cysteine protease cathepsin F is the most abundant protein produced by fourth-stage larvae (L4) of the nematode Teladorsagia circumcincta, while its transcript is only detectable in L4 and adults. T. circumcincta cathepsin F is a recently evolved cysteine protease that does not fall clearly into either of the cathepsin L or F subfamilies. This protein exhibits characteristics of both cathepsins F and L, and its phylogenetic relationship to its closest homologs is distant, including proteins of closely related nematodes of the same subfamily.

KEYWORDS: Cysteine protease, cathepsin, Teladorsagia circumcincta, bioinformatics, homology modeling, gastrointestinal nematode

RECEIVED: August 12, 2020. ACCEPTED: September 2, 2020.

TYPE: Original Research

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by a grant from La Trobe University.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Sarah Sloan, AgriBio Centre for AgriBioscience, Department of Animal, Plant and Soil Sciences, School of Life Sciences, La Trobe University, 5 Ring Road, Bundoora, Victoria 3086, Australia. Email: s.sloan@latrobe.edu.au

Introduction

Nematoda are an ancient group and most terrestrial plants and larger animals are associated with at least 1 parasitic nematode.¹ Although fossilization of various nematode species has occurred throughout history, as early as 135 million years ago (mya),² there are no records of *Teladorsagia circumcincta* fossils and, as such, an estimate of their time of divergence is difficult to calculate. However, the earliest bovids appeared 20 mya³ and since T. circumcincta is seen in sheep and goats but not cattle, and there is a closely related nematode in cattle, Ostertagia ostertagi, we can reasonably assume their divergence from related nematode species occurred within this time-frame, likely alongside the host. As such, T. circumcincta is a relatively recently evolved nematode when compared to the long history of the nematoda phylum, and some of its mechanisms of immune evasion within the host subsequently relatively recent as well. Some secretory proteins of T. circumcincta are likely to have adapted to suit the specific modern host.

Teladorsagia circumcincta is the most important parasitic nematode of sheep in cool temperate regions worldwide.⁴ Eggs are passed in feces and develop into infective third-stage larvae (L3) on pasture, which are ingested by the host. L3 exsheath in the rumen, enter the lumen of the abomasal glands and molt to become fourth-stage larvae (L4) and establish within the host. L4 then molt into sexually mature adults which feed and breed on the mucosal surface of the abomasa.⁵ Teladorsagiosis can result in reduced production, decreased animal welfare, parasitic gastroenteritis, poor growth performance, and weight loss.6 Several methods are used to control T. circumcincta infection including anthelmintic treatment, nutritional supplementation, vaccination, selective breeding, and pasture management. These methods are already used with varying success,⁷⁻¹¹ however, T. circumcincta is developing drug resistance.4,12 A vaccine against Teladorsagiosis was developed by Nisbet et al,¹³ which uses larval antigens that are targets of IgA antibodies from immune sheep. IgA plays a crucial role in protection against T. circumcincta.14 One of these antigen targets was cathepsin F, a cysteine protease.

Parasitic cysteine proteases are involved in parasite stage transition, invasion of host tissues, nutrient uptake, and immune evasion.¹⁵⁻¹⁸ The papain family of clan CA is the largest and most abundant family of cysteine proteases,¹⁹ and consists of 2 major subfamilies; cathepsin B and cathepsin L. Cathepsin B is characterized by the presence of an occluding peptide loop,²⁰ while the cathepsin L subfamily is characterized by the presence of an ERF/WNIN motif, and comprises cathepsins L, V, K, S, W, F, and H.^{19,21} Cathepsin L pro-regions are about 100 residues long with 2 conserved motifs; ERF/WNIN and GNFD.²¹ Cathepsin F has a longer pro-region, up to 250 residues, and, in general, is composed of 2 domains: an N-terminal cystatin-like domain and a C-terminal peptide similar to the cathepsin L pro-region.²² However, the motifs in the C-terminal peptide are different in cathepsin F; a highly conserved ERFNAQ motif replaces the ERF/WNIN motif, and adjacent to this is an E/DxGTA motif, which was identified as a pro-region feature of cathepsins F and W, but not cathepsin L.23 E/DxGTA has been suggested to act together with ERFNAQ as a scaffold for the pro-region, maintaining the inhibitory specificity of its α-helical structure.¹⁹ The GNFD motif seen in cathepsin L is also present in cathepsin F and was previously identified as critical in intermolecular processing, and as the site of initial cleavage in Fasciola hepatica cathepsin L.^{24,25} Cathepsin L of the parasitic trematode F. hepatica has been shown to cleave and digest host IgG.26

Cathepsin F in *T. circumcincta* is the most abundant protein produced by L4, and its transcript is detectable only in fourthstage larvae (L4) and adult nematode stages, not in pre-parasitic stages.^{23,27} The functions of cathepsin F in T. circumcincta and



Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage). its host interactions are yet to be determined, however, this protease may have a similar role to that of cathepsin L in *F. hepatica*, which is to digest IgG as a protective mechanism against the host immune response.²⁶ Secretion of cathepsin F could be digesting host IgA for protection against the host immune response or as a source of protein to aid in the development of the nematode.

This study aimed to use bioinformatic analyses to evaluate the protein sequence, structure, and gene of cathepsin F, and to find homologs with other nematode species to determine its phylogenetic history and potential functional role. This information is intended to inform the design of further functional studies.

Materials and Methods

Gene analysis and assembly

The mRNA sequence for T. circumcincta cathepsin F (GenBank accession no. DQ133568) and its translated protein sequence (Tci-CF-1, GenBank accession no. ABA01328) were obtained from The European Bioinformatics Institute (EBI). The mRNA sequence for Tci-CF-1 was used as a query for BLASTn against the draft T. circumcincta genome in WormBase ParaSite²⁸ (BioProject: PRJNA72569, Taxonomy ID: 45464) and against all nematode genomes in the database to identify matching genes. Predicted gene exons that matched Tci-CF-1 (>80% identity and E-value threshold of 4.5E-18) were extracted, translated into protein sequences, and aligned with Tci-CF-1 using CLC Genomics Workbench Version 9 (CLC, Qiagen) and default parameters. CLC alignments use a progressive alignment algorithm. The most closely aligned reading frame was selected and exons were re-ordered to align with Tci-CF-1.

Variants of Tci-CF-1 were constructed using PacBio RS and Illumina HiSeq 2500 sequence reads obtained from the Sequence Read Archive (BioProject PRJEB7676).²⁹ Sequence reads were converted to FASTq format using SRA Toolkit 2.8.2 (https://github.com/ncbi/sra-tools). The quality of reads was checked with FastQC 0.11.6 (https://github.com/sandrews/FastQC) using default parameters. Low-quality Illumina reads were trimmed with Trimmomatic 0.32 (https:// github.com/timflutre/trimmomatic). CLC was used to assemble the Illumina reads using Tci-CF-1 as the reference (variants 1 and 2). SPAdes 3.10.1 (https://github.com/ablab/ spades) was used for de novo assembly of combined PacBio and Illumina reads. BLASTn of the mRNA sequence for Tci-CF-1 against the SPAdes de novo assembly in CLC resulting in contigs containing matches with >75% identity were extracted and used to assemble the Tci-CF-1 gene (variant 3).

Phylogenetic analysis and multiple sequence alignments

Tci-CF-1 was used as a query for BLASTp analysis against EBI and the National Center for Biotechnology Information databases. The selected sequences had an e-value threshold of 1e-50. Whole sequences were extracted into CLC and duplicates were removed. One *T. circumcincta* papain family cysteine protease (GenBank accession no. PIO64159) matched and was excluded following further analysis which identified a misassembly of its corresponding gene resulting in incorrect protein sequence formation; the exons were in the wrong order (data not shown). Pairwise comparisons of the 172 total sequences were conducted, and the top 9 sequences with the highest percent identity (% ID) to Tci-CF-1 were selected for multiple sequence alignment and pairwise comparison.

A multiple sequence alignment of Tci-CF-1 and the 9 closest homologs identified in the BLASTp analysis was carried out using CLC under default parameters. An alignment between Tci-CF-1, the 3 variants identified in this study and the possible *T. circumcincta* cathepsin F polymorphism discussed in Nisbet et al⁷ was carried out using CLC. The translated protein sequences were annotated using information from previous studies.^{21,23,24,30} Prediction of N-glycosylation sites was conducted via NetNGlyc 1.0 Server.³¹

Phylogenetic analysis was performed using SplitsTree5 5.0.0_alpha.³² The multiple sequence alignment of Tci-CF-1 and the 9 closest homologs was used as the original input and consisted of 8 taxa and 10 protein character sequences of length 475. The Neighbor Joining method³³ and Tree Embedder method³⁴ were used (default options) to obtain a rooted tree drawing, and bootstrap values calculated following 1000 replications. The Uncorrected_P method³⁵ was used (default options) to obtain a 10 × 10 distance matrix. The Neighbor Net method³⁶ and The Splits Network Algorithm method³⁷ were used (default options) to obtain a splits network. The Splits Network Algorithm method³⁷ was used (ReticulateNetwork splits transformation, default options) to obtain a reticulate splits network.

Prediction of secondary and tertiary structure

Homology modeling of Tci-CF-1, variants 1, 2, and 3, and the Nisbet et al¹³ possible variant protein sequences (Figure 1), as well as the pro-regions of human cathepsins L (UniProtKB accession no. P07711) and F (UniProtKB accession no. Q9UBX1) was conducted using Protein Homology/analogY Recognition Engine V 2.0 (Phyre², Kelley et al³⁸) under the intensive modeling mode. The predicted protein structures were analyzed using UCSF ChimeraX.³⁹

Results

Gene analysis and assembly

Tci-CF-1 matched several exons from the draft genome assembly with exons of predicted genes TELCIR_20397, _06733, _06734, _14223, and _19209 (BioProject: PRJNA72569). Sequence _20397 has a forward orientation, while all remaining sequences are on the reverse strand. Tci-CF-1 matched at exons 2 and 3 of TELCIR_20397, matched at exons 1-3, 5 and 6 of TELCIR_06733, matched at

Tci-CF-1	MSLLFLLLIP	HLFAATVKQQ	YSGGVKPLTE	LRTDLIDKKT	KGSIEFARLG	QHISPKDFGA
Nisbet et al. (2013) polymorph Variant 1	• • • • • • • • • •	• • • • • • • • • • •		• • • • • • • • • •	•••'1'••••• *	
Variant 2			D			N
Variant 3			D			· · · · · <u>N</u> . · · ·
T : of (INUIDEODIDD					
ICI-CF-1 Nichot at al. (2012) notymorph	WNHFTSFIER	HDKVIRNESE	ALKREGIEKR	NLEIIRSAQE	NDKGTALIGI	NQFADLSPEE
Variant 1	• • • • • • • • • • •	····	• • • • • • • • • • •	• • • • • • • • • • •	D	• • • • • • • • • • • •
Variant 2				.SR	V	TR
Variant 3		K				
Tci-CF-1	FKKTHLPHTW	KQPDHPNRIV	DLAAEGVDPK	EPLPESFDWR	EHGAVTKVKT	EGHCAA
Nisbet et al. (2013) polymorph	••••A•	•••••••	• • • • • • • • • • •	• • • • • • • • • • •	• • • • • • • • • • •	• • • • • • • • • • •
Variant 2	•••••••••	••••••••••	••••••••••	••••••••••	•••••	•••••••••
Variant 3						
Tci-CF-1	SVTGNIEGQW	FLAKKKLVSL	SAQQLLDCDV	VDEGCNGGFP	LDAYKEIVRM	GGLEPEDKYP
Nisbet et al. (2013) polymorph						<u></u>
Variant 1					CR	· · · · S · · · · ·
Variant 2	• • • • • • • • • • •	• • • • • • • • • • •	• • • • • • • • • • •	• • • • • • • • • • •	CR	· · · · S
Variant 5	• • • • • • • • • • •	• • • • • • • • • • •	• • • • • • • • • • •	• • • • • • • • • • •	• • • • • • • • • • •	•••••••••••••••••••••••••••••••••••••••
				IN CONTRACTOR	DIGIGINIDD	TODUKOGUOD
Tci-CF-1	YEAKAEQCRL	VPSDIAVYIN	GSVELPHDEE	KMRAWLVKKG	PISIGITVDD	IQFIKGGVSR
Nisbet et al. (2013) polymorph	• • • • • • • • • • •	• • • • • • • • • • •	• • • • • • • • • • •	• • • • • • • • • • •	• • • • • • • • • •	• • • • • • • • • • •
Variant 2	••••••••••	••••••••••	••••••••••	••••••••••••••••••••••••••••••••••••••	••••••••••	••••••••••
Variant 3						
Tci-CF-1	PTTCRLSSMI	HGALLVGYGV	EKNIPYWIIK	NSWGPNWGED	GYYRMVRGEN	ACRINRFPTS
Nisbet et al. (2013) polymorph	KPY	.		.		
Variant 1	· · · · · P. · · ·				\dots \underbrace{I}_{\pm} \dots \ldots	
Variant 2	•••••••••••••••••••••••••••••••••••••••	• • • • • • • • • • •	• • • • • • • • • • •	• • • • • • • • • • •	••••	••G•••••
Variant S	•••••••••••••••••••••••••••••••••••••••	• • • • • • • • • • •	• • • • • • • • • • •	• • • • • • • • • • •	• • • • • • • • • • •	• • • • • • • • • • •
Tci-CF-1	AVVI					
Nisbet <i>et al.</i> (2013) polymorph						
Variant 1	• • • •					
Variant 2						
Variant 3	• • • •					
re 1 Multiple sequence alignme	ont of Teladorsadia	circumcincta secr	eted cathensin F (1	ci-CE-1 GenBank	accession no ARA	01328) the

Figure 1. Multiple sequence alignment of *Teladorsagia circumcincta* secreted cathepsin F (Tci-CF-1, GenBank accession no. ABA01328), the polymorphism discussed in Nisbet et al¹³ and variants 1, 2 and 3 identified in this study. Conserved residues indicated by a dot; stop codon by an asterisk; ERFNAQ, E/DxGTA, GxNxFxD and GCNGG motifs by square, cross, circle, and diamond, respectively; catalytic triad residues by a downward arrow; predicted N-glycosylation sites by a right-facing arrow; polymorphisms of interest in boxes.

exons 1, 2, and 4 of TELCIR_06734, matched at exons 8, 2-6, in that order, as well as some areas in the introns of TELCIR_14223, and matched at exons 1-3, as well as areas up- and down-stream, of TELCIR_19209 (Supplemental Data 1).

None of the genes in the databases encoded the complete cathepsin F sequence. Between exons and between genes were several large sections of incomplete assembly or gaps which may contain the missing regions. TELCIR_14223 was an almost-perfect match to Tci-CF-1 for the exons available. TELCIR_06733 and _06734 are adjacent in the genome. Together the separate sequences can encode the entire Tci-CF-1 sequence with 68.7% similarity, and all motifs are conserved (Supplemental Data 2). The matching of exons from different loci may be a consequence of partial sequencing of tandemly repeated genes or errors in the assembly of the genome.

The consensus sequences of the 3 variants all have high similarity to Tci-CF-1. Variant 3 gave a nucleotide and amino acid sequence similarity of 98.6% (1080/1095 nucleotides, 359/364 amino acids) to Tci-CF-1. Variants 1 and 2 gave nucleotide similarities of 98% (1073/1095 nucleotides) and 97.3% (1066/1095 nucleotides), respectively, and amino acid sequence similarities of 97.3% (354/364 amino acids) and

96.4% (351/364 amino acids), respectively. An alignment of Tci-CF-1, the 3 variants identified in this study and the Nisbet et al¹³ variant show the signal peptide, catalytic triad and predicted N-glycosylation sites are conserved, as well as most of the motifs (Figure 1). The presence of a stop codon in variant 1 implies that there are at least 1 functional gene and 1 pseudogene in the *T. circumcincta* genome, but a sequencing error cannot be ruled out.

The gene structure of cathepsin F was determined from the PacBio and Illumina combined reads (variant 3), and has a minimum length of 9583 bp over 10 exons and 9 introns. Exons 1-10 are 119 bp, 90 bp, 70 bp, 90 bp, 145 bp, 160 bp, 85 bp, 94 bp, 178 bp, and 64 bp in length, respectively, resulting in a 364 amino acid protein, and the exon phase class is 0-2-2-0-0-1-2-0-1-2-0-1. Introns 1-9 are >937 bp, 173 bp, 1427 bp, 998 bp, 338 bp, 439 bp, 1882 bp, 705 bp, and >1589 bp in length, respectively (Figure 2).

Phylogenetic analysis

The 9 complete sequences with the highest % ID on January 13, 2020 to Tci-CF-1 were *Diploscapter pachys* hypothetical protein WR25_25536 (GenBank accession no. PAV60527,



Figure 2. *Teladorsagia circumcincta* cathepsin F variant 3 gene exon/intron structure, constructed using PacBio and Illumina reads in SPAdes and CLC. Exons indicated by solid black arrows, introns indicated by white segments, gaps in contigs indicated by a break. The positions of intron-exon junctions and phase classes are denoted by diamonds.

Table 1. Pairwise comparison of complete homologous protein sequences to Tci-CF-1 by sequence identity (%) and distance.

	TCI-CF-1	DP-HP-1	HC-PI-1	HC-PI-2	DV-CF-1	ACA-HP	SV-UP	ACO-UP	DP-HP-2	AC-PFCP
Tci-CF-1		48.51	47.13	46.71	45.18	44.28	44.54	43.81	43.74	42.71
Dp-HP-1	0.54		47.63	47.41	47.17	45.73	51.65	47.65	66.37	48.71
Hc-PI-1	0.48	0.34		98.7	73.06	75.65	45.26	63.79	59.62	48.81
Hc-PI-2	0.49	0.34	0.01		73.28	75.65	45.04	63.36	59.41	48.6
Dv-CF-1	0.53	0.35	0.3	0.3		71.98	43.7	60.78	58.35	45.47
Aca-HP	0.56	0.39	0.26	0.26	0.3		43.33	75.44	57.08	45.79
Sv-UP	0.48	0.34	0.23	0.23	0.27	0.29		42.82	38.19	67.73
Aco-UP	0.57	0.35	0.31	0.32	0.35	0.16	0.3		54.49	41.23
Dp-HP-2	0.61	0.03	0.46	0.46	0.47	0.48	0.42	0.48		39.43
Ac-PFCP	0.4	0.38	0.24	0.25	0.29	0.28	0.12	0.34	0.47	

Upper quadrant: sequence identity (%); lower quadrant: distance; Tci-CF-1: *Teladorsagia circumcincta* secreted cathepsin F (GenBank accession no. ABA01328); Dp-HP-1: *Diploscapter pachys* hypothetical protein WR25_25536 (GenBank accession no. PAV60527); Hc-PI-1: *Haemonchus contortus* proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain containing protein (GenBank accession no. CDJ88889); Hc-PI-2: *H. contortus* proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain containing protein (GenBank accession no. CDJ88889); Hc-PI-2: *H. contortus* proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain containing protein (GenBank accession no. CDJ88889); Hc-PI-2: *H. contortus* proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain containing protein (GenBank accession no. CDJ88889); Hc-PI-2: *H. contortus* proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain containing protein (GenBank accession no. CDJ88889); Hc-PI-2: *H. contortus* proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain containing protein (GenBank accession no. CDJ88889); Hc-PI-2: *H. contortus* proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain containing protein (GenBank accession no. CDJ88889); Hc-PI-2: *H. contortus* proteinase inhibitor I26 and proteinase inhibitor I29 and peptidase C1A domain containing protein (GenBank accession no. CDJ8262); Dv-CF-1: *Dictyocaulus viviparus* cathepsin F1 (GenBank accession no. AFM37363); Ac-PFCP: *Ancylostoma ceylanicum* papain family cysteine protease (GenBank accession no. EPB70524); Aca-HP: *Angiostrongylus cantonensis* hypothetical protein Angca_010213 (GenBank accession no. KAE9418773); Sv-UP: *Strongylus vulgaris* unnamed protein product (GenBank accession no. VDM61191); Dp-HP-2: *D. pachys* hypothetical protein WR25_24125 (GenBank accession no. PA67875).

309 residues), Haemonchus contortus proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain-containing protein (GenBank accession no. CDJ92562, 463 residues), H. contortus proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain-containing protein (GenBank accession no. CDJ88889, 463 residues), Dictyocaulus viviparus cathepsin F1 (GenBank accession no. AFM37363, 459 residues), Ancylostoma ceylanicum papain family cysteine protease (GenBank accession no. EPB70524, 287 residues), Angiostrongylus cantonensis hypothetical protein Angca_010 213 (GenBank accession no. KAE9418773, 456 residues), Strongylus vulgaris unnamed protein product (GenBank accession no. VDM81154, 264 residues), Angiostrongylus costaricensis unnamed protein product (GenBank accession no. VDM61191, 405 residues), and *D. pachys* hypothetical protein WR25_24125 (GenBank accession no. PAV67875, 452 residues) (Table 1).

Pairwise comparisons showed that Tci-CF-1 (GenBank accession no. DQ133568, 364 residues) has highest %ID with

a *Diploscapter pachys* hypothetical protein at 48.51% ID, respectively. The remaining 8 closest homologs ranged from 42.71-47.13% ID (Table 1).

Phylogenetic analysis showed a rooted tree drawing with 18 nodes and 17 edges, and 4 separate clades (Figure 3). The Neighbor-Net splits network had 35 nodes and 48 edges (Figure 4A) and complements the clades in the tree. The reticulate splits network had 37 nodes and 50 edges (Figure 4B). Both split networks resulted in 20 cyclic splits. Table 1 indicates the distances between taxa calculated in the multiple sequence alignment.

Sequence analysis

The nucleotide sequence implies a translated protein of 364 amino acids. The first 14 amino acids are the signal sequence while amino acids 15 to 150 form the pro-region. The mature protein goes from amino acids 151 (Glutamic acid) to 364



Figure 3. Phylogenetic tree of Tci-CF-1 and its 9 closest homologs. Tci-CF-1: Teladorsagia circumcincta secreted cathepsin F (GenBank accession no. ABA01328); Dp-HP-1: Diploscapter pachys hypothetical protein WR25_25536 (GenBank accession no. PAV60527); Hc-PI-1: Haemonchus contortus proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain-containing protein (GenBank accession no. CDJ88889); Hc-PI-2: H. contortus proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain-containing protein (GenBank accession no. CDJ92562); Dv-CF-1: Dictyocaulus viviparus cathepsin F1 (GenBank accession no. AFM37363); Ac-PFCP: Ancylostoma ceylanicum papain family cysteine protease (GenBank accession no. EPB70524); Aca-HP: Angiostrongylus cantonensis hypothetical protein Angca_010213 (GenBank accession no. KAE9418773); Sv-UP: Strongylus vulgaris unnamed protein product (GenBank accession no. VDM81154); Aco-UP: Angiostrongylus costaricensis unnamed protein product (GenBank accession no. VDM61191); Dp-HP-2: D. pachys hypothetical protein WR25_24125 (GenBank accession no. PAV67875). Scale-bar indicates branch lengths and bootstrap values are indicated following 1000 replications.

(Leucine). Protein sequence analysis of Tci-CF-1 revealed several conserved features that are typical of cathepsin F proteins. Tci-CF-1 has a hydrophobic signal sequence (residues 1-14) with a signal cleavage site between residues 14 and 15 (Figure 5). The pro-region contains cathepsin F-like motifs such as $E_{80}R_{84}F_{88}N_{91}A_{95}Q_{99}$, in which alanine (A_{95}) has been substituted for isoleucine (I_{95}), $E/D_{102}xG_{104}T_{105}A_{106}$, and $G_{109}xN_{111}xF_{113}xD_{115}$ (Figure 6). The predicted N-terminal of the mature processed protein is located at residue E_{151} (Figure 5). The mature protein contains the conserved structural motif $G_{214}C_{215}N_{216}G_{217}G_{218}$ as well as the catalytic triad active site residues C_{177} , H_{311} , and N_{331} . Two predicted N-glycosylation sites were found at residues $N_{77}E_{78}S_{79}$ and $N_{260}G_{261}S_{262}$ (Figure 6).

Multiple sequence alignments

The multiple sequence alignment of the most similar sequences indicates that 6 out of the 9 sequences contain a region of amino acids in the pro-region (~97 amino acids) that is not

present in the pro-region of Tci-CF-1, corresponding to a cystatin domain. The catalytic triad residues and motifs are mostly conserved amongst all homologous sequences. However, *Strongylus vulgaris* and *Ancylostoma ceylanicum* are missing the histidine and asparagine residues of the catalytic triad (Figure 7).

Secondary and tertiary structure

Homology modeling using Phyre² revealed that the secondary structure of Tci-CF-1 comprised 37% alpha-helices, and 15% beta-strands. The tertiary structure has 100% confidence in homology to cysteine proteases, and 53% ID with cysteine protease folds from the papain-like family. The predicted structure for residues 1-60 has low confidence, while the predicted structure for residues 61-364 has high confidence (Supplemental Data 3). In the tertiary structure model, the 3 amino acids that form the catalytic triad come together to form the active site, and the inhibitor domain appears to block access to the catalytic triad (Figure 8). Polymorphisms were present on the periphery of the structure at positions 30 ($E_{30} > D_{30}$), 56 ($K_{56} > N_{56}$), 76 ($R_{76} > K_{76}$), 235 ($P_{235} > S_{235}$), and 306 ($L_{306} > P_{306}$) (Figures 1 and 8).

Homology modeling of Tci-CF-1 illustrates structural similarity to x-ray crystallographic human cathepsin L and F structures.^{40,41} The mature domains appear structurally similar (Figure 9) despite the aminoacid sequences being quite different between proteins, and the pro-region of human cathepsin L is more structurally similar to Tci-CF-1 (Figure 9B and C) than human cathepsin F (Figure 9A).

Discussion

Bioinformatic analysis of Cathepsin F of T. circumcincta has assembled a putative gene encoding the protein, described the gene structure, discovered several potential polymorphisms, failed to identify any close homologs and predicted the structure of the protein. Most cysteine proteases are synthesized as an inactive precursor. Cathepsins, a family within the cysteine proteases, maintain typical features including an N-terminal hydrophobic signal peptide sequence, a pro-peptide domain, and a mature domain containing the active site.²³ The active site consists of a catalytic triad of cysteine, histidine, and asparagine residues.^{42,43} Cathepsin pro-peptide inhibitor domains contain an α -helical structure that prevents access to the active site and proteolytic cleavage is needed to remove the inhibitor and activate the zymogen⁴⁴ (Figure 8). Conserved motifs of cysteine proteases are critical for characterization of the different sub-families.

When cathepsin F of *T. circumcincta* was first identified, its closest homologues were hypothetical proteins of *Caenorhabditis* briggsae and *Caenorhabditis elegans*.²³ The highest % identity in the databases on April 2020 is with a *D. pachys* hypothetical protein and a *H. contortus* proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain-containing



Figure 4. Phylogenetic networks of Tci-CF-1 and its 9 closest homologues; Tci-CF-1: *Teladorsagia circumcincta* secreted cathepsin F (GenBank accession no. ABA01328); Dp-HP-1: *Diploscapter pachys* hypothetical protein WR25_25536 (GenBank accession no. PAV60527); Hc-PI-1: *Haemonchus contortus* proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain-containing protein (GenBank accession no. CDJ88889); Hc-PI-2: *H. contortus* proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain-containing protein (GenBank accession no. CDJ88889); Hc-PI-2: *H. contortus* proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain-containing protein (GenBank accession no. CDJ92562); Dv-CF-1: *Dictyocaulus viviparus* cathepsin F1 (GenBank accession no. AFM37363); Ac-PFCP: *Ancylostoma ceylanicum* papain family cysteine protease (GenBank accession no. EPB70524); Aca-HP: *Angiostrongylus cantonensis* hypothetical protein Angca_010213 (GenBank accession no. KAE9418773); Sv-UP: *Strongylus vulgaris* unnamed protein product (GenBank accession no. VDM81154); Aco-UP: *Angiostrongylus costaricensis* unnamed protein product (GenBank accession no. VDM61191); Dp-HP-2: *D. pachys* hypothetical protein WR25_24125 (GenBank accession no. PAV67875). (A) Neighbor-Net split network. (B) Reticulate network. Scale-bar indicates the distance of the edges.

protein at 48.51% and 47.13%. The similarity between these proteins is not particularly high, and molecules from species other than T. circumcincta appear more closely related to one another, ranging from 38.19% to 98.7% identity (Table 1). Closely related species are expected to have more similar proteins. For example, tropomyosin from T. circumcincta and Trichostrongylus colubriformis has been shown to differ by only 1 amino acid,⁴⁵ both nematodes belong to the order Strongylida and are in the same clade V. Cathepsin L of Fasciola hepatica has been shown to digest IgG.²⁶ If Tci-CF-1 has a similar role, it could suppress the immune response given the important role that antibody responses play in resistance to this parasite.⁴⁶ The absence of close homologs suggests that this protein has emerged relatively recently in evolutionary history. Since there are no fossil records of T. circumcincta specifically, an estimate of its divergence from relatives must be made. The earliest bovids appeared 20 mya,³ and since T. circumcincta is seen in sheep and goats but not cattle, we can safely assume their divergence from related nematode species occurred following or alongside bovid divergence. Chilton et al⁴⁷⁻⁴⁹ have shown the evolutionary relationships between T. circumcincta (sheep and goats), O. ostertagi (cattle), and H. contortus (sheep and goats), and that they are very closely related. There was no similar protein detected in this study between T. circumcincta and O. ostertagi, the more closely related of these 3 species, but there was a protein

detected from *H. contortus* (Table 1), a slightly more distant species but still within the same Trichostrongylidae clade and with the same host species. Whether the gene for cathepsin F in *T. circumcincta* was inherited from an ancestor or developed later is unknown, but it has diverged enough from relatives to be a distinct protein.

One of the methods in which a new gene arises is by exon shuffling, however, there was no evidence of ancestral exons, as there is no 1-1 class phase observed in the cathepsin F gene (Figure 2).50 Gene assembly demonstrated that the Tci-CF-1 gene is composed of 10 exons spanning a minimum length of 9.5 kbp. The SNPs with transcribed substitutions, resulting in the 5 polymorphisms identified in this study, are located on the periphery of the protein structure (Figure 8A). Of particular interest are the 2 polymorphisms located in the mature protein (proline (P_{235}) to serine (S_{235}), and leucine (L_{306}) to P_{306}), as these are likely to affect mature protein function and/or recognition. The substitution of P_{235} to S_{235} is conservative as proline and serine are both small amino acids and serine is commonly present within tight turns on protein surfaces because its hydroxyl oxygen can form a hydrogen bond with the protein backbone and mimic proline. 51 The location of the $\mathrm{L}_{\mathrm{306}}$ in Tci-CF-1 indicates that it may be influencing the catalytic triad in some way. Leucine is a hydrophobic amino acid and its position on the outside of the protein structure suggests that it

AT	GTC	ТСТ	TTT	GTT	CCT	GCT	TCT	CAT	ссс	ACA	TCT	ATT	TGC	CGC	TAC	TGT	AAA	GCA	GCA	ATA	СТС	AGG	AGG	TGT	CAA	ACC	GTT	GACA
M	S	L	L	F	L	L	L	I	Ρ	Η	L	F	A	А	Т	V	Κ	Q	Q	Y	S	G	G	V	Κ	Ρ	L	Т
CA	አ ጥጥ	ccc	መእሮ	CC 7	ጥጥጥ	$C \lambda T$	CCA	C 7 7		CNC	C 7 7	лсс	ירייר	$C \lambda T$	CCA	C = T			CCT	тсс	ሞሮ እ	707	Слт		TCC	<u>, , ,</u>	лсл	CTTC
E	L	R	T	D	L	IAD	D	K	GAA K	GAC T	K	G	S	I	E	F	A	R.	L	G	0	H	I	S	P	K K	D	F
																					~							
GG	TGC	ATG	GAA	TCA	TTT	CAC	CAG	CTT	CAT	TGA	AAG	GCA	TGA	CAA	GGT	СТА	CAG	AAA	CGA	GAG	CGA	AGC	TCT	GAA.	ACG.	ATT	TGG	GATC
G	А	W	Ν	Η	F	Т	S	F	Ι	Е	R	Η	D	Κ	V	Y	R	N	E		<u></u> ∎E	A	L	K	R	F	G	I
														×		x	×	x)				
ΤT	CAA	GAG	AAA	TCT	'CGA	GAT	AAT	TCG	СТС	TGC	GCA	GGA	AAA	CGA	TAA	GGG	AAC	AGC	TAT	TTA	CGG	AAT	CAA	TCA	GTT	TGC	TGA	TCTT
F	K	R	Ν	L	Ε	Ι	Ι	R	S	A	Q	Ε	Ν	D	K	G	Т	A	Ι	Y	G	Ι	Ν	Q	F	А	D	L
TC	ACC	GGA	gga	ATT	'CAA	AAA	.GAC	TCA	ССТ	GCC	GCA	CAC	ATG	GAA	ACA	GCC	TGA	TCA	TCC.	AAA	CCG	AAT	CGT	GGA	CTT.	AGC	CGC	AGAA
S	P	E	E	F	K	K	Т	Н	L	Р	Н	Т	W	K	Q	P	D	Н	P	Ν	R	I	V	D	L	A	A	E
					\checkmark																							
GG	GGT	GGA	TCC	GAA	.GGA	GCC	ACT	GCC	GGA	ATC	GTT	CGA	TTG	GAG	AGA	ACA	TGG	TGC	AGT	GAC	AAA	AGT	'GAA	AAC'	TGA	AGG	TCA	CTGT
G	V	D	Ρ	K	Ε	Ρ	L	Ρ	Ε	S	F	D	W	R	Ε	Η	G	А	V	Т	Κ	V	Κ	Т	Ε	G	Η	С
GC	ACC	CTC	CTTC	ccr	יתיית	ጥጥሮ	ጥርጥ	CAC	ACC	222	ጥፚጥ	тсa	ACC	CCA	стс	ርጥጥ	CCT	יידכר		מממ	CDD	аст	יידכיד	מימי	CCT	CTTC	ccc	
A	A		W	A	F	S	V	T	G	N	I	E	G G	Q	W	F	L	A	K	K	K	L	V	S	L	S	A	Q
		-																										
CA	GCT	ССТ	CGA	TTG	TGA	TGT	TGT	TGA	TGA	GGG	ATG	TAA	.CGG	▼ TGG	ATT	TCC	тст	'TGA	CGC	TTA	CAA	AGA	AAT	CGT	TCG	AAT	GGG	CGGC
Q	L	L	D	С	D	V	V	D	Ε	G	С	Ν	G	G	F	Ρ	L	D	А	Y	Κ	Ε	I	V	R	Μ	G	G
mm	CCA	» cc		7 ~ 7	<u> </u>	CULY	maa	CILIN	007	» c c	<u> </u>	ccc		CCA	CIIIC	maa	COM	mem	000	3 00 0	CC 7	<u>т</u> л т		mem	ጠጠእ		<u> </u>	cccc
L	E E	P P	aga E	aga D	K	ιgia Υ	P	Y	CGA E	AGC A	K	JJJJ. A	.aga E	GCA 0	C	R	L	V	P	S	D	IAI	A	V	Y	IAI	N	G
1		-	2	2		-	-	-	2				-	×	Ũ		-	·	-	2	2	-		•	-	-		~~~~
TC	AGT	CGA	GCT	ACC	ACA	TGA	TGA	AGA.	AAA	AAT	GAG	GGC	ATG	GCT	AGT	GAA	GAA	GGG	GCC	GAT	ATC	GAT	'AGG	TAT	CAC	CGT	AGA	TGAC
<u>S</u>		Ε	L	Ρ	Η	D	Ε	Ε	K	М	R	A	W	L	V	K	K	G	Ρ	Ι	S	Ι	G	Ι	Т	V	D	D
AT	ACA	GTT	ста	TAA	AGG	CGG	CGT	TTC	TCG	TCC	GAC	TAC	CTG	TAG	ACT	ATC	TTC	TAT	GAT	TCA	TGG	CGC	TCT	CCT	GGT	CGG	ATA	CGGT
I	Q	F	Y	K	G	G	V	S	R	Ρ	Т	Т	С	R	L	S	S	М	Ι	H	G	A	L	L	V	G	Y	G
0.0	001	~ ~ ~		— ~ —		0.003	000	C A D		~ ~ ~	~ 7 7			~~~					7 0 7	001	паа	7 m 7		a a	0 7 0			maga
GT V	CGA E	GAA K	GAA N	TAT	P	GTA V	UTG W	GAT	TAT	AAA K	.GAA Ni	S	GTG W	999 G	P	CAA N	TTG W	טטטי G	AGA E	GGA D	TGG G	ATA Y	Y I'I'A	.CAG	GAT	GG.T. V	R	G
v	-	11		-	-	-	••	-	-			D	••	0	-	1.	••	0	ц		0	-	-	1.		v	1.	5
GA	GAA	CGC	TTG	TCG	CAT	AAA	CAG	ATT	ССС	CAC	GTC	AGC	TGT	TGT	CCT	ATA	A											
Ε	Ν	А	С	R	I	Ν	R	F	Ρ	Т	S	А	V	V	L	*	-											
re 5.	Ann	otate	d Tel	ador	sagia	a circ	umci	ncta	secre	eted	cathe	epsin	Fse	quen	ce (G	aenB	ank a	acces	sion	no. E	Q13	3568	3) wit	n ann	otatio	ons.	Signa	al seque

Figure 5. Annotated *Teladorsagia circumcincta* secreted cathepsin F sequence (GenBank accession no. DQ133568) with annotations. Signal sequence is <u>underlined</u>; predicted N-glycosylation sites marked with <u>squiggle underline</u>; N-terminal amino acid of mature protein indicated by black arrow; catalytic triad active site residues highlighted black; ERFNAQ, E/DxGTA, GxNxFxD and GCNGG motifs indicated by a square, cross, circle and diamond, respectively.

may be involved in substrate recognition.⁵¹ The substitution of leucine to proline is interesting due to the ability of proline to introduce kinks into the sequence. The substitution of leucine to proline may change the way the catalytic triad interacts with host molecules.

The pro-region of Tci-CF-1 does not have typical cathepsin F characteristics. A typical cathepsin F has a long pro-region (up to 250 residues) and in general, the pro-peptide of cathepsin F is composed of 2 domains; an N-terminal cystatin-like domain and a C-terminal peptide similar to the cathepsin L pro-region.²² In Tci-CF-1, homology modeling indicates that the cystatin-like domain is missing (Figures 7 and 9), resulting in a much shorter pro-region, and it is more similar to a typical cathepsin L pro-region. *Clonorchis sinensis* cathepsin F (GenBank accession no. AF093243) is also missing the cystatin-like domain in the pro-region of its cathepsin F protein.¹⁹

Typical cathepsin F pro-regions also contain a highly conserved ERFNAQ motif, in place of the cathepsin L ERF/ WNIN motif .43 Adjacent to the ERFNAQ motif is an E/ DxGTA motif, which was identified as a pro-region feature of cathepsins F and W, but not cathepsin L.23 These 2 motifs together are thought to act as a scaffold of the pro-region and maintain the inhibitory function of the α -helical structures of cathepsins F and W.19 Redmond et al23 demonstrated that Schistosoma mansoni cathepsin L (GenBank accession no. AAC46485) contains the ERFNAQ motif, and not the expected ERF/WNIN motif, which is consistent with the Tci-CF-1 in the current study.43,52 In addition, S. mansoni cathepsin L does not contain a cystatin-like domain, which is consistent with Cathepsin L, but also with Tci-CF-1 in the current study (Figure 9). Despite the conservation of the cathepsin F/W E/DxGTA motif in Tci-CF-1, the lack of a cystatin-like



Figure 6. Homology model of *Teladorsagia circumcincta* secreted cathepsin F (GenBank accession no. ABA01328). (A) Ribbon structure showing alpha helices, beta sheets, catalytic triad residues (C, H, N), predicted N-glycosylation sites, ERFNAQ, E/DxGTA, GCNGG, and GxNxFxD motifs. (B) Magnified ERFNAQ, E/DxGTA, GxNxFxD motifs and pro-region N-glycosylation site. (C) Magnified catalytic triad and GCNGG motif. (D) Magnified mature domain N-glycosylation site. Sidechains of residues of interest labeled and bold.

domain in Tci-CF-1 coupled with the presence of the ERFNAQ motif in *S. mansoni* cathepsin L, illustrates that the distinctions between cathepsins F and L may not be as clear as often assumed.

Substitutions at specific residues between cathepsins F and L may provide insights into Tci-CF-1. The substitution of alanine (A) to isoleucine (I) in the ERFNAQ motif of Tci-CF-1 places this motif halfway between ERFNAQ (cathepsin F) and ERF/WNIN (cathepsin L). Both alanine and isoleucine can be readily substituted for one another due to their small size, and non-reactive sidechains. The ERF/WNIN and ERFNAQ motifs are known to form α-helical structures,⁵¹ however, isoleucine is known to be restricted in its conformations, and finds it difficult to form an α -helical structure. Similarly, asparagine (N) can be readily substituted for glutamine (Q) as both these residues can be substituted by polar amino acids, are similar in structure, and are frequently involved in protein active or binding sites, of which the ERFNAQ and ERFNIN motifs inhibit.51 Whether Tci-CF-1 has evolved from ERFNAQ or ERFNIN to ERFNIQ remains unknown. Phylogenetic analysis of our top 10 similar sequences does not provide insights either.

Tci-CF-1 is quite isolated from the rest of the proteins as seen in Figures 2 and 3 and illustrates that although these homologous proteins are closest by sequence %ID, they are closer to one-another than Tci-CF-1. Phylogenetic networks show the possible relationships in a dataset. Taxa are represented by nodes and their evolutionary relationships are represented by edges.³² Recombination, hybridization, gene conversion and gene transfer all lead to phylogenetic relationships that cannot be adequately modeled by a single tree. Even when the underlying history is treelike, sampling error and parallel evolution may make it difficult to establish a single, accurate phylogenetic tree.³² Parallel edges are used to represent the splits of the taxa, instead of single branches of a tree.³⁶ Split networks often contain nodes that do not represent ancestral species and, therefore, can only provide a suggestive representation of evolutionary history. Two split network methods were applied in this study to portray the relationship between Tci-CF-1 and its 9 closest homologs: The Neighbor-Net split network and the Reticulate Network.

Neighbor-Net split networks use multiple sequence alignment distance calculations to construct a circular collection of weighted splits and can represent conflicting signals in the data, whether they arise from sampling error or genuine recombinations³⁶ and show the uncertainty of the phylogenetic history. The more tree-like the network, the more confidence that the tree constructed is an accurate representation of the phylogenetic history with the data used. Figure 4A shows that the groupings observed in the network are complementary to the clades in the phylogenetic tree (Figure 3). Tci-CF-1 is distinctly isolated from the other groupings. The Neighbor-Net gives confidence that although there are several alternative phylogenetic tree outputs possible, they will follow roughly the same paths, with the different clades consistently grouping together (Figure 4A).

Reticulate Networks illustrate evolutionary histories, and their splits are a result of reticulate events such as hybridization, horizontal gene transfer, or recombination. The internal nodes represent hypothetical ancestral species, and nodes with 2 or

		20		40		60		80	
Tci-CE-1	MSLIFILLP			1			VK00		30
Dp-HP-1	M								1
Hc-PI-1	M H VW	FVFLIPHLFA	ASIKEKESGE	VNPLTELHKE	LAQKAVDDMN	AKSNDLYRWD	VYTIDQVHQQ	MHNGIQYKMH	74
Hc-PI-2	M H VW	FVFLIPQLFA	ASIKEKESGE	VNPLTELHKE	LAQKAVDDMN	AKSNDLYRWD	VYTIDQVHQQ	MHNGIQYKMH	74
Dv-CF-1	MI-FW	SLLLIPQLFA	GTLKEEANPS	AQLYQE	LAQKAVHKYN	EKSNDLYMWD	IFSIDNVHQQ	IHNGIHYHMK	70
Aca-HP	MMALW	LLYLLPQLIA	KS	PNSLNELHQE	LAQKAVDNYN	EKSNDLYRWD	IYSIDDVQQK	MHNGIQYNMK	67 10
Aco-UP	M ADV -	SPISP	SS	KPGC LL YKNE	LAOKAVDNYN	EKSNDI YRWD	LYSIDDVOOK		61
Dp-HP-2	MRVFVLSLT-	-LFVISSLHA	AKMLGAMSEA	EDFDEQQKNE	YAQKAVDKYN	AESNDLFKWE	LDSVDSVNKQ	IVTGIKVDME	78
Ac-PFCP	M I PW	VILVIPQILA	GTIKESAE-K	PDPLTEVHQ-	VDSL -				36
		100		120		140		160	
Tci-CF-1	LRTDLIDKKT	KGS						- E F A R L G Q H I	53
Dp-HP-1							I D	TEELRKSHRV	13
Hc-PI-1	LTLVQTDCRK	TGSELTRTDC	KHTDRLKKCT	AQGEHRAWEN	YENIEIRHCE	QPFKRSKR	-MAKTSKHSV	E - YERLSRH I	150
Hc-PI-2	LTLVQTDCRK	TGTELTRTDC	KHTDRLKKCT	AQGEHRAWEN	YENIEIRHCE	QPFKRSKR	-MAKTSKHSV	E - YKRLSRH I	150
Dv-CF-1	LTLVETTCPK	SVNQLSLNNC	KHTDRLKKCT	AEGEHHAWEN	YENIEVHHCT	EPYQRTKR	- MLKTVLSPI	Q-HKEFLHYI	146
Aca-HP		AVDSV	KQIDRLQKCI	AEGEHRAWEK	YENVEIRHCA	EPFRRIKR	-MIRMMHIPS		143
Aco-UP		MVS	VS	SPPGLSSESE	SESIPLEECI	FENSY		- 0K3KV3KH1	101
Dp-HP-2	FTMVQTDCTK	QAK I		SHRAWDK	FEKVELKGCE	EPFQRSKKDR	VRKRLAHMVD	TEELRKSHRV	139
Ac-PFCP						- P F K R T	P V	D - TAKLGRH I	52
		180	_	200		220	•	240	
Tci-CE-1	SPKDEGAWNH	FTSFIERHDK				GTALYGINGE	ADISPEEKK		133
Dp-HP-1	KRKDIIPWNQ	FMDFIDEHDK	NYSNKREILK	KFKIFKTNLK	AIKMYQQKEH	GTAVYGITQF	ADLTPAQFKK	IYLPYQWEEP	93
Hc-PI-1	KPRDYVAWNQ	FTDFVDRHDK	LYSSKREALK	RFRIFKRNLK	AIRTWQENEK	GTAVYGITQF	SDLSPEEFKK	IYLPYTWEQP	230
Hc-PI-2	KPRDYVAWNQ	FTDFVDRHDK	LYSSKREALK	RFRIFKRNLK	AIRTWQENEK	GTAVYGITQF	SDLSPEEFKK	IYLPYTWEQP	230
Dv-CF-1	KPKDYVAWNQ	FVDFMGRHEK	VYNSKHDTLK	RFRVFKRNLK	AIRSWQEKEE	GTAVYGITQF	SDLTPEEFKK	IYLPYIWDEP	226
Aca-HP	KPKDFVAWNQ	FTDFLERYNK	AHSDKHETMK	RFRIFKRNLK	AIRLWQEREE	GTAVYGVTQF	SDLTPDEFKK	IYLPYIWDQP	223
SV-UP	KPKDYVAWNQ	FSDFLDRHEK	VYDSKRDILK	RFRIFKRNLK		GTAVYGTTQF	SDLIPEEFKK		114
Dn-HP-2	KRKDIJPWNO		NYSNKREILK	KEKIEKTNIK	AIKMYOOKEH	GTAVYGITOF	ADI TPAOFKK		219
Ac-PFCP	KPKDYVAWNQ	FTEFIDRHEK	TYNGKRDTLK	RERIEKRNIK	AIRTWOENEE	GTAVYGITOF	SDITPEEKK	IYLPYIWEEP	132
					, <u></u>	OTAVIOTIQ			102
		260		280		300	ODE IT EET KK	320	102
Tci-CE-1			PESEDWREHG						213
Tci-CF-1 Dp-HP-1	DHPNR I VDL A	AEGVDPKEPL AEGVYVEDAL	PESFDWREHG	280 I AVTKVKTEGH AVTSVKNOGM	CAACWAFSVT CGSCWAFSTT	GN I EGQWFLA GN VEGOWFLA	KKKLVSLSAQ	320 I QLLDCDVVDE ELVDCDKLDQ	213 173
Tci-CF-1 Dp-HP-1 Hc-PI-1	DHPNR I VDL A I I PNR VAN LE I YPNR I AD LA	260 I AEGVDPKEPL AEGVYVEDAL AEGISLDDPL	PESFDWREHG PETFDWREHG PDSFDWRTHG	280 I AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT	GN I EGQWFLA GN VEGQWFLA GN I EGQWYLA	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ	320 I QLLDCDVVDE ELVDCDKLDQ ELVDCDIIDQ	213 173 310
Tci-CF-1 Dp-HP-1 Hc-PI-1 Hc-PI-2	DHPNR I VDLA I I PNR VANLE I YPNR I ADLA I YPNR I ADLA	260 I AEGVDPKEPL AEGVYVEDAL AEGISLDDPL AEGISLDDPL	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG	280 I AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGN	CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT	GN I EGQWFLA GN VEGQWFLA GN I EGQWYLA GN I EGQWYLA	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLLSLSEQ	QLLDCDVVDE ELVDCDKLDQ ELVDCDIIDQ ELVDCDIIDQ	213 173 310 310
Tci-CF-1 Dp-HP-1 Hc-PI-1 Hc-PI-2 Dv-CF-1	DHPNR I VDL A I I PNR VAN LE I YPNR I ADL A I YPNR I ADL A I VPNR WVDL T	AEGVDPKEPL AEGVYVEDAL AEGISLDDPL AEGISLDDPL AEGVHLNETL	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG	280 I AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGF	¢ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT	GN I EGQWFLA GN VEGQWFLA GN I EGQWYLA GN I EGQWYLA GN I EGQWFLA	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLLSLSEQ KKKLVSLSEQ	320 I QLLDCDVVDE ELVDCDKLDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDKVDD	213 173 310 310 306
Tci-CF-1 Dp-HP-1 Hc-PI-1 Hc-PI-2 Dv-CF-1 Aca-HP	DHPNR I VDLA I I PNR VANLE I YPNR I ADLA I YPNR I ADLA I VPNRMVDLT I FPNK I VDLE	AEGVDPKEPL AEGVVVEDAL AEGISLDDPL AEGISLDDPL AEGVHLNETL AEGINTTDDY	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG PDSFDWRKHG	2000 I AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGF AVTEVKNQGF	¢ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT	3000 I GN I EGQWFLA GN VEGQWFLA GN I EGQWYLA GN I EGQWFLA GN I EGQWFLA	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLUSLSEQ KKKLVSLSEQ	ALLDCDVVDE ELVDCDKLDQ ELVDCDIDQ ELVDCDIDQ ELVDCDIDQ ELVDCDIDQ	213 173 310 310 306 303
Tci-CF-1 Dp-HP-1 Hc-PI-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP	DHPNR I VDLA I I PNR VANLE I YPNR I ADLA I YPNR I ADLA I VPNRMVDLT I FPNK I VDLE I YPNK I ADLE	260 I AEGVDPKEPL AEGVYVEDAL AEGISLDDPL AEGISLDDPL AEGVHLNETL AEGINTTDDY AEGINEL	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG PDSFDWRKHG PESFDWRFKG	2000 I AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGN AVTDVKNQGS AVTEVKNQGS AVTEVKNQGS	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT	3000 I GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLUSLSEQ KKKLVSLSEQ KKKLVSLSEQ	3200 EL VDCDK LDQ EL VDCDK LDQ EL VDCD I IDQ EL VDCD I IDQ EL VDCD I IDQ EL VDCD I VDQ	213 173 310 310 306 303 191
Tci-CF-1 Dp-HP-1 Hc-Pl-1 Hc-Pl-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Do-HP-2	DHPNRIVDLA I IPNRVANLE I YPNRIADLA I YPNRIADLA I VPNRMVDLT I FPNKIVDLE I YPNKIADLE I I PNRMANLE	260 I AEGVDPKEPL AEGVSLDDPL AEGISLDDPL AEGINTDDL AEGINTTDDY AEGINTTDDY AEGINVTDDY	PESFDWREHG PETFDWREHG PDSFDWRTHG PESFDWRDHG PDSFDWRKHG PESFDWRPKG PDSFDWREHG PETFDWREHG	2000 I AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGF AVTEVKNQGS AVTEVKNQGS AVTEVKNQGN AVTEVKNQGN	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT	3000 I GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ	ALLOCOVVDE ALLOCOVVDE ELVDCDKLDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIVDQ	213 173 310 310 306 303 191 252 299
Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Ac-PFCP	DHPNR I VDLA I IPNR VANLE I YPNR I ADLA I YPNR I ADLA I VPNRMVDLT I FPNK I VDLE I FPNK I VDLE I FPNK I VDLE I IPNRMANLE I YPNK I ADLE	260 I AEGVDPKEPL AEGVJVEDAL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINEL AEGINVTDDY AEGIVVEDAL AEGIDVKEDL	PESFDWREHG PETFDWREHG PDSFDWRTHG PESFDWRDHG PDSFDWRKHG PESFDWREHG PESFDWREHG PESFDWRKKG	2000 I AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGF AVTEVKNQGN AVTEVKNQGN AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT	3000 I GN I EGQWFLA GN V EGQWFLA GN I EGQWFLA	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ	3200 I QLLDCDVVDE ELVDCDKLDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIVDQ ELVDCDIVDQ ELVDCDKLDQ ELVDCDVVDQ	213 173 310 306 303 191 252 299 212
Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Ac-PFCP	DHPNR I VDLA I IPNR VANLE I YPNR I ADLA I YPNR I ADLA I VPNRMVDLT I FPNK I VDLE I FPNK I VDLE I FPNK I VDLE I IPNRMANLE I YPNK I ADLE	260 I AEGVDPKEPL AEGVSLDDPL AEGISLDDPL AEGISLDDPL AEGINTDDY AEGINTDDY AEGINVTDDY AEGIDVKEDL AEGIDVKEDL 340	PESFDWREHG PETFDWREHG PDSFDWRTHG PESFDWRDHG PDSFDWRKHG PESFDWREHG PESFDWREHG PESFDWRKKG	2200 I AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGF AVTEVKNQGN AVTDVKNQGS AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM 300	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT	3000 I GN I EGQWFLA GN I EGQWFLA 380	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ	3200 I QLLDCDVVDE ELVDCDKLDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIVDQ ELVDCDIVDQ ELVDCDIDQ ELVDCDIVDQ A000	213 173 310 306 303 191 252 299 212
Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Ac-PFCP	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRWVDLT IFPNKIVDLE IYPNKIADLE IFPNKIVDLE IFPNKIADLE	260 I AEGVDPKEPL AEGVYVEDAL AEGISLDDPL AEGISLDDPL AEGINTDDY AEGINTDDY AEGINVTDDY AEGIDVKEDL 340 I XKELVPMCCI	PESFDWREHG PETFDWREHG PDSFDWRTHG PESFDWRDHG PDSFDWRKG PDSFDWRKG PESFDWREHG PESFDWRKKG	2000 1 AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGF AVTEVKNQGN AVTDVKNQGS AVTEVKNQGN AVTGVKNQGM AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM 3000 1	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT	3000 I GN I EGQWFLA GN I EGQWFLA SN I EGQWFLA GN I EGQWFLA SN I EGQWFLA	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ	3200 I QLLDCDVVDE ELVDCDKLDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIVDQ ELVDCDIVDQ ELVDCDIVDQ ELVDCDIVDQ I UVDCDKLDQ ELVDCDVVDQ I	213 173 310 306 303 191 252 299 212
Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Ac-PFCP Tci-CF-1 Dp-HP-1	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRWVDLT IFPNKIVDLE IYPNKIADLE IFPNKIVDLE IIPNRMANLE IYPNKIADLE	260 I AEGVDPKEPL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINEL AEGINVTDDY AEGVYVEDAL AEGIDVKEDL 340 I YKEIVRMGGL YKEILBIGGI	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRCHG PESFDWREHG PESFDWREHG PESFDWREHG PESFDWREKG EPEDKYPYEA	2000 1 AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGN AVTDVKNQGF AVTEVKNQGN AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM AVTEVKNQGN AVTEVKNQGM AVTEVKNQM AVTEVKNQM AVTEVKNQM AVTEVKNQM AVTEVKNQM AVTEVKNQM AVTEVKNQM AVTEVKNQM AVTEVKNQM AVTEVKNQM AVTEVKNQM AVTEVKNM	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT	3000 I GN I EGQWFLA GN V EGQWFLA GN V EGQWFLA SN V EGQWFLA I EL PHDEEKMR EL PHDEK MR	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ AWLVKKGPIS SWLVQKGPIS	3200 I QLLDCDVVDE ELVDCDKLDQ ELVDCDHDQ ELVDCDHDQ ELVDCDHDQ ELVDCDHVDQ ELVDCDHVDQ ELVDCDKLDQ ELVDCDKLDQ IGHVDDIGF IGHVDDIGF	213 173 310 306 303 191 252 299 212 293 253
Tci-CF-1 Dp-HP-1 Hc-Pl-1 Hc-Pl-2 Dv-CF-1 Aca-HP Sv-UP Dp-HP-2 Ac-PFCP Tci-CF-1 Dp-HP-1 Hc-Pl-1	DHPNRIVDLA I IPNRVANLE I YPNRIADLA I YPNRIADLA I VPNRWVDLT I FPNKIVDLE I YPNKIADLE I FPNKIVDLE I IPNRMANLE I YPNKIADLE GCNGGFPLDA GCEGGLPSNA	260 I AEGVDPKEPL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINTTDDY AEGINVTDDY AEGINVTDDY AEGIDVKEDL 31 JKEIVRMGGL YKEIIRLGGL YKEIIRMGGL	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG PDSFDWREHG PDSFDWREHG PETFDWREHG PESFDWRKKG EPEDKYPYEA ESEDKYPYDA EPEDAYPYDG	2000 1 AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGS AVTEVKNQGS AVTEVKNQGN AVTDVKNQGS AVTEVKNQGM AVTGVKNQGM 300 1 KAEQCRLVPS RGESCHVAKP KGETCHIVRS	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT DIAVYINGSV DIAVYINDSL EIAVYINDSL	30100 A GN I EGQWFLA GN VEGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA 301 ELPHDEEKMR ELPHDEEKMR ELPHDEESMK	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ AWLVKKGPIS SWLVQKGPIS AWLVKKGPIS	3200 I QLLDCDVVDE ELVDCDKLDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIVDQ ELVDCDIVDQ ELVDCDIVDQ ELVDCDKLDQ ELVDCDKLDQ IGITVDDIQF IGINATPLQF IGINANPLQF	213 173 310 306 303 191 252 299 212 293 253 390
Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Aco-PFCP Tci-CF-1 Dp-HP-1 Hc-PI-2	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRMVDLT IFPNKIVDLE IFPNKIVDLE IIPNRMANLE IYPNKIADLE GCNGGFPLDA GCCGGLPSQA GCNGGLPSQA	260 I AEGVDPKEPL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINTTDDY AEGINVTDDY AEGINVTDDY AEGIDVKEDL I YKEIVRMGGL YKEIIRLGGL YKEIIRMGGL	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRPKG PDSFDWREHG PESFDWREHG PESFDWREKG PESFDWRKKG EPEDKYPYEA ESEDKYPYEA EPEDAYPYDG EPEDAYPYDG	2000 1 AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGM AVTEVKNQGM AVTEVKNQGN 300 1 KAEQCRLVPS RGESCHVAKP KGETCHIVRS KGETCHIVRS	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT DIAVYINGSV DISVYINDSL EIAVYINDSM EIAVYINDSM	3000 A GN I EGQWFLA GN VEGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN VEGQWFLA GN VEGQWFLA BN VEGQWFLA S80 1 ELPHDEEKMR ELPHDEEKMR ELPHDEESMK	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KHQLVSLSEQ AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS	3200 I QLLDCDVVDE ELVDCDKLDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIVDQ ELVDCDIVDQ ELVDCDIVDQ ELVDCDKLDQ ELVDCDKLDQ IGITVDDIQF IGINANPLQF IGINANPLQF	 213 173 310 306 303 191 252 299 212 293 253 390 390
Tci-CF-1 Dp-HP-1 Hc-PI-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Ac-PFCP Tci-CF-1 Dp-HP-1 Hc-PI-1 Hc-PI-2 Dv-CF-1	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRMVDLT IFPNKIVDLE IYPNKIADLE IFPNKIVDLE IIPNRMANLE QCNGGFPLDA GCNGGFPLDA GCNGGLPSQA GCEGGLPSQA	260 I AEGVDPKEPL AEGVYVEDAL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINTTDDY AEGINVTDDY AEGINVTDDY AEGIDVKEDL 340 I YKEIVRMGGL YKEIIRMGGL YKEIMRMGGL	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG PDSFDWREHG PESFDWREHG PETFDWREHG PESFDWRKKG EPEDKYPYEA ESEDKYPYDA EPEDAYPYDG ETESAYPYDG	2200 1 AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGN AVTDVKNQGF AVTEVKNQGN AVTDVKNQGS AVTEVKNQGN AVTEVKNQGN 1 KAEQCRLVPS RGESCHVAKP KGETCHIVRS KGETCHIVRS KGETCHINRT	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT DIAVYINGSV DISVYINDSL EIAVYINDSV EFAVYINDSV	30100 GN I EGQWFLA GN VEGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN VEGQWFLA GN VEGQWFLA GN VEGQWFLA S800 1 EL PHDEEKMR EL PHDEESMK EL PHDEESMK	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KHQLVSLSEQ AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS	3200 I QLLDCDVVDE ELVDCDKLDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIVDQ ELVDCDIVDQ ELVDCDKLDQ ELVDCDKLDQ IGITVDDIQF IGINANPLQF IGINANPLQF IGINANPLQF	213 173 310 306 303 191 252 299 212 293 253 390 386
Tci-CF-1 Dp-HP-1 Hc-PI-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Aca-HP Tci-CF-1 Dp-HP-2 Hc-PI-1 Hc-PI-2 Dv-CF-1 Aca-HP	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRMVDLT IFPNKIVDLE IYPNKIADLE IFPNKIVDLE IIPNRMANLE IFPNKIADLE CONGGFPLDA GCEGGLPSQA GCEGGLPSQA GCEGGLPSQA	260 I AEGVDPKEPL AEGVIVEDAL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINTTDDY AEGINVTDDY AEGIDVKEDL AEGIDVKEDL 340 I YKEIVRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRDHG PESFDWRDHG PESFDWREHG PESFDWREHG PESFDWRKKG EPEDKYPYEA ESEDKYPYDA EPEDAYPYDG ETESAYPYDG ESESDYPYDG	2000 1 AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGF AVTEVKNQGN AVTDVKNQGS AVTEVKNQGN AVTEVKNQGM AVTEVKNQGN 3600 1 KAEQCRLVPS RGESCHVAKP KGETCHIVRS KGETCHIVRS RGEECHINRT KGEKCHIVRS	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT DIAVYINGSV DISVYINDSL EIAVYINDSV EIAVYINDSV EIAVYINDSV	GN I EGQWFLA GN VEGQWFLA GN VEGQWFLA GN VEGQWFLA SN I EGQWFLA CN I EGQWFLA CN I EGQWFLA CN I EGQWFLA CN I EGQWFLA SN I EGQWFLA CN I EGQWFLA SN I EGQWFLA CN I EGQWFLA SN I EGQWFLA CN I EGQWFLA SN I EGQ	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKLVSLSEQ KHQLVSLSEQ AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS	I GINANPLQF IGINANPLQF	213 173 310 306 303 191 252 299 212 293 253 390 386 383
Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Dp-HP-2 Ac-PFCP Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Sv-UP	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRWVDLT IFPNKIVDLE IYPNKIADLE IFPNKIVDLE IIPNRMANLE IYPNKIADLE CONGGEPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA	260 I AEGVDPKEPL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINEL AEGINVTDDY AEGVYVEDAL AEGIDVKEDL 340 YKEIVRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRCHG PESFDWREHG PESFDWREHG PESFDWREKG EPEDKYPYEA ESEDKYPYDA EPEDAYPYDG ETESAYPYDG ESESDYPYDG ESESDYPYDG	200 200 1 AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTDVKNQGF AVTEVKNQGN AVTDVKNQGS AVTEVKNQGM AVTGVKNQGM AVTGVKNQGM AVTGVKNQGM AVTEVKNQGN AVTEV	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT DIAVYINGSV DISVYINDSU EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSV	3000 GN I EGQWFLA GN VEGQWFLA GN I EGQWYLA GN I EGQWFLA GN I EGQWFLA BL PHDEEKMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKLVSLSEQ KKLVSLSEQ AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS	3200 I QLLDCDVVDE ELVDCDKLDQ ELVDCDKLDQ ELVDCDKVDD ELVDCDKVDD ELVDCDKVDQ ELVDCDKLDQ ELVDCDKLDQ ELVDCDKLDQ IGITVDDIQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF	213 173 310 306 303 191 252 299 212 293 253 390 386 383 263 383
Tci-CF-1 Dp-HP-1 Hc-PI-1 Dv-CF-1 Aca-HP Dy-CF-1 Aco-UP Dp-HP-2 Aco-PFCP Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Sv-UP	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRWVDLT IFPNKIVDLE IYPNKIADLE IFPNKIVDLE IIPNRKADLE GONGGFPLDA GCEGGLPSQA GCEGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA	260 I AEGVDPKEPL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINEL AEGINVTDDY AEGVYVEDAL AEGIDVKEDL JKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG PESFDWREHG PESFDWREHG PESFDWREHG PESFDWREKG EPEDKYPYEA ESEDKYPYDA EPEDAYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG	2000 1 AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGS AVTEVKNQGS AVTEVKNQGM AVTDVKNQGS AVTEVKNQGM AVTGVKNQGM AVTGVKNQGM AVTEVKNQGN 1 KAEQCRLVPS RGESCHVAKP KGETCHIVRS RGEECHIVRS RGETCHLVRS KGETCHLVRS KGETCHLVRS KGECCHUAKS	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT DIAVYINGSV DISVYINDSL EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSV	GN I EGQWFLA GN I EGQWFLA EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS	3200 I QLLDCDVVDE ELVDCDKLDQ ELVDCDKLDQ ELVDCDIDQ ELVDCDIDQ ELVDCDIDQ ELVDCDKLDQ ELVDCDKLDQ ELVDCDKLDQ IGITVDDIQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF	213 173 310 306 303 191 252 299 212 293 253 390 386 383 263 332 270
Тсі-СF-1 Dp-HP-1 Hс-PI-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Ac-PFCP Tci-CF-1 Hс-PI-1 Hс-PI-2 Dv-CF-1 Hc-PI-2 Sv-UP Sv-UP Aco-UP Dp-HP-2 Aco-UP	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRMVDLT IFPNKIVDLE IFPNKIVDLE IIPNKIADLE IFPNKIADLE GCNGGFPLDA GCEGGLPSAA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA	260 I AEGVDPKEPL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINTTDDY AEGINTTDDY AEGIVVEDAL AEGIDVKEDL JKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG PDSFDWREHG PDSFDWREHG PETFDWREHG PESFDWRKKG EPEDKYPYEA ESEDKYPYDA EPEDAYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDA ESEDRYPYDA ESEDRYPYDA	200 200 1 AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGN AVTDVKNQGS AVTEVKNQGM AVTDVKNQGS AVTEVKNQGM AVTGVKNQGM AVTEVKNQGM 300 1 KAEQCRLVPS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGESCHVAKP KGESCHVAKP	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT DIAVYINDSV DIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSL EIAVYINDSL EIAVYINDSL EIAVYINDSL	GN I EGQWFLA GN I EGQWFLA BN I EGQWFLA GN I EGQWFLA BL PHDEEKMR EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS	3200 I QLLDCDVVDE ELVDCDKLDQ ELVDCDKLDQ ELVDCDKVDD ELVDCDKVDD ELVDCDKVDQ ELVDCDKVDQ ELVDCDKLDQ ELVDCDKLDQ IGITVDDIQF IGINANPLQF IGINANPLQF IGINANPLQF IGINATPLQF IGINATPLQF IGINATPLQF IGINATPLQF IGINATPLQF IGINATPLQF IGINATPLQF	213 173 310 306 303 191 252 299 212 293 253 390 390 386 383 263 332 2379 2386
Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Aco-PFCP Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Ac-PFCP	DHPNRIVDLA I IPNRVANLE I YPNRIADLA I YPNRIADLA I VPNRWVDLT I FPNKIVDLE I FPNKIVDLE I IPNRKADLE I PNRKIADLE GCNGGFPLDA GCCGGLPSAA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA	260 I AEGVDPKEPL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINTTDDY AEGINVTDQU AEGINVTDQU AEGIN	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG PDSFDWREHG PDSFDWREHG PESFDWREHG PESFDWREKG EPEDAYPYDA EPEDAYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESEDRYPYDA EPEDAYPYDG	200 1 AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGN AVTDVKNQGS AVTEVKNQGN AVTDVKNQGS AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM AVTEVKNQGN 300 1 KAEQCRLVPS RGESCHVAKP KGETCHIVRS RGEECHIVRS RGEECHIVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT DIAVYINGSV DISVYINDSL EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSL EIAVYINDSL EIAVYINDSL	GN I EGQWFLA GN I EGQWFLA BN I EGQWFLA GN I EGQWFLA BL PHDEEKMR EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS SWLVQKGPIS AWLVKKGPIS	320 I QLLDCDVVDE ELVDCDKLDQ ELVDCDKLDQ ELVDCDKVDD ELVDCDKVDQ ELVDCDIDQ ELVDCDKDQ ELVDCDKDQ IGITVDDIQF IGINANPLQF IGINANPLQF IGINATPLQF IGINATPLQF IGINATPLQF IGINATPLQF IGINATPLQF IGINATPLQF IGINATPLQF IGINATPLQF IGINATPLQF	213 173 310 306 303 191 252 299 212 293 253 390 386 383 263 332 379 286
Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Ac-PFCP Tci-CF-1 Dp-HP-2 Ac-PFCP Tci-CF-1 Hc-PI-1 Hc-PI-2 Dv-CF-1 Hc-PI-2 Aca-HP Sv-UP Aco-UP Dp-HP-2 Ac-PFCP	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRMVDLT IFPNKIVDLE IIPNKIVDLE IIPNKIADLE GCNGGFPLDA GCEGGLPSOA GCNGGLPSOA GCNGGLPSOA GCNGGLPSOA GCNGGLPSOA GCNGGLPSOA	260 I AEGVDPKEPL AEGVYVEDAL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINTTDDY AEGINVTDY AEGINVT	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG PDSFDWREHG PESFDWREHG PETFDWREHG PESFDWRKKG EPEDKYPYDA EPEDAYPYDG ETESAYPYDG ESESDYPYDG ESESDYPYDG ESEDRYPYDA EPEDAYPYDG	200 200 1 AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGS AVTEVKNQGN AVTDVKNQGS AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM AVTEVKNQGN 300 1 KAEQCRLVPS RGESCHVAKP KGETCHIVRS RGEECHIVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS 404 440 440 440 440 440 440 44	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT DIAVYINGSV DISVYINDSU EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSU EIAVYINDSU EIAVYINDSL EIAVYINDSL	GN I EGQWFLA GN I EGQWFLA ELPHDEEKMR ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS SWLVQKGPIS AWLVKKGPIS	3200 I QL LDCD VVDE EL VDCDK LDQ EL VDCD I IDQ EL VDCD VDQ I GI NANPLQF I GI NANPLQF I GI NANPLQF I GI NANPLQF I GI NATPLQF I GI NATPLQF	213 173 310 306 303 191 252 299 212 293 253 390 386 383 263 332 379 286
Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Dp-HP-2 Ac-PFCP Tci-CF-1 Dp-HP-2 Ac-PFCP Tci-CF-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Aco-PFCP	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRMVDLT IFPNKIVDLE IYPNKIADLE IFPNKIVDLE IIPNRMANLE IFPNKIADLE CONGGFPLDA GCNGGFPLDA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA	260 I AEGVDPKEPL AEGVYVEDAL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINTTDDY AEGINVTDDY AEGINVTDDY AEGINVTDDY AEGINVTDDY AEGINVTDDY AEGINVTDDY AEGINVTDDY AEGINVTDDY AEGINVTDDY AEGINVTDU I YKEIVRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG PDSFDWREHG PESFDWREHG PESFDWREHG PESFDWREKG EPEDKYPYDA EPEDAYPYDG ESESDYPYDG ESESDYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG	200 200 1 AVTKVKTEGH AVTSVKNQGM AVTSVKNQGN AVTAVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN 300 1 KAEQCRLVPS RGESCHVAKP KGETCHIVRS RGEECHIVRS RGECCHIVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS 140 140 140 140 140 140 140 140	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT DIAVYINGSV DISVYINDSL EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSL SWGPNWGEDG	3000 GN I EGQWFLA GN VEGQWFLA GN I EGQWFLA GN I EGQWFLA EL PHDEEKMR EL PHDEESMK EL PHDESMK EL PHDESMK CANANA COMPANY COMP	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKLVSLSEQ KHQLVSLSEQ AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS SWLVQKGPIS AWLVKKGPIS CRINR-FPTS	3200 I 3200 I ALLDCDVVDE ELVDCDKUDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIVDQ ELVDCDIVDQ ELVDCDVVDQ ELVDCDVVDQ IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINATPLQF IGNTATPLQF IGNTATPLQF	213 173 310 306 303 191 252 299 212 293 253 390 386 383 263 332 379 286
Тсі-СF-1 Dp-HP-1 Hс-PI-2 Dv-CF-1 Aca-HP Sv-UP Dp-HP-2 Ac-PFCP Tсі-CF-1 Dp-HP-1 Aca-HP Sv-UP Sv-UP Tсі-CF-1 Dp-HP-1 Hс-PI-2	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRWVDLT IFPNKIVDLE IYPNKIADLE IFPNKIVDLE IIPNRMANLE IYPNKIADLE GCNGGFPLDA GCEGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA	260 I AEGVDPKEPL AEGVYVEDAL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGVYVEDAL AEGINVTDDY AEGVYVEDAL AEGIDVKEDL 300 310 311 314 314 314 314 314 314 314	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRCHG PESFDWREHG PESFDWREHG PESFDWREHG PESFDWREKG EPEDKYPYDA ESEDKYPYDA EPEDAYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG GALLVGYGVE GVLIVGYGSE	AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGN AVTAVKNQGN AVTDVKNQGF AVTEVKNQGN AVTDVKNQGS AVTEVKNQGN AVTEVKNVX AVTEVKX AVTEXX AVTEVKX AVTEVKX AVTEXX		3000 GN I EGQWFLA GN VEGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN VEGQWFLA GN VEGQWFLA GN VEGQWFLA GN VEGQWFLA GN VEGQWFLA GN VEGQWFLA S800 1 ELPHDEEKMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK A00 2 HODESMK A00 1 1 1 1 1 1 1 1 1 1 1 1 1	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKLVSLSEQ KKLVSLSEQ AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS	3200 I 3200 I ALLDCDVVDE ELVDCDKLDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIIDQ ELVDCDIVDQ ELVDCDKLDQ ELVDCDKLDQ ELVDCDKLDQ IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINATPLQF	213 310 310 306 303 191 252 299 212 293 253 390 386 383 363 332 379 286
Тсі-СF-1 Dp-HP-1 Hс-PI-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Aco-FFCP Tсі-CF-1 Dp-HP-1 Hс-PI-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Aco-PFCP	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRWVDLT IFPNKIVDLE IYPNKIADLE IFPNKIVDLE IIPNKADLE IFPNKIADLE CONGGFPLDA GCNGGFPLDA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA CONGCLPSQA CONGCLPSQA CONGCLPSQA CONGCLPSQA CONGCLPSQA CONGCLPSQA CONGCLPSQA CONGCLPSQA CONGCLPSQA CONGCLPSQA CONGCLPSQA CONGCLPSQA CONGCLPSQA CONGCLPSQA CONGCN CONC CONC CONC CONC CONC CONC CO	260 I AEGVDPKEPL AEGVYVEDAL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINVTDDY AEGINVTDDY AEGVYVEDAL AEGIDVKEDL ³⁴⁰ 1 YKEIVRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIRMGGL YKEIRMK YKEIRMGL YKEIRMK YKEIRMGL YKEIRMGL YKEIRMK YKEIRMK YKEIRMK YKEIRMK YKEIRMK YKEI YKEIRMK YKEIRMK YKEI Y	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG PESFDWREHG PESFDWREHG PESFDWREHG PESFDWREHG PESFDWREKG EPEDKYPYEA ESEDKYPYDA EPEDAYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA	200 200 200 200 200 200 200 200	↓ CAACWAFSVT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT CGSCWAFSTT DIAVYINGSV DISVYINDSL EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSV EIAVYINDSL SWGPNWGEDG SWGTKWGEEG SWGTKWGEEG	GN I EGQWFLA GN I EGQWFLA EL PHDEESMK EL PHDEESMK A460 I YYRMVRGENA YYR YDE YYRL YRGKNV YYRL YRGKNV	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS CRINR-FPTS CGVHEM-ATS CGVHEM-ATS	3200 I QLLDCDVVDE ELVDCDKLDQ ELVDCDKLDQ ELVDCDIDQ ELVDCDIDQ ELVDCDIVDQ ELVDCDKLDQ ELVDCDKLDQ ELVDCDKLDQ IGITVDDIQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINATPLQF	213 1173 310 306 303 191 252 299 212 293 253 390 386 383 332 379 286
Тсі-СF-1 Dp-HP-1 Hс-PI-1 Ac9-IP Dy-CF-1 Ac9-IP Dp-HP-2 Ac9-PFCP Tсі-CF-1 Dp-HP-1 Hс-PI-1 Hс-PI-2 Ac9-FFCP Тсі-CF-1 Dp-HP-2 Ac9-UP Dp-HP-2 Ac9-UP Dp-HP-2 Ac9-UP Dp-HP-2 Ac9-UP Dp-HP-2 Ac9-UP Dp-HP-2 Dy-CF-1 Hc-PI-1 Hc-PI-2 Dy-CF-1 Dy-HP-1 Hc-PI-2 Dy-CF-1 Dy-HP-1 Hc-PI-2 Dy-CF-1 Dy-HP-2 Dy-Dy-2 Dy-D	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRWVDLT IFPNKIVDLE IYPNKIADLE IFPNKIVDLE IIPNRMANLE IYPNKIADLE CONGGFPLDA GCEGGLPSQA GCEGGLPSQA GCNGGLPSQA	260 I AEGVDPKEPL AEGVYVEDAL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINTTDDY AEGINTTDDY AEGVYVEDAL AEGIDVKEDL 340 1 YKEIVRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGGL YKEIIRMGL YKEIIRMGL YKEIIRMGL YKEIRMGL HTCRLSSMIH IFCAPYMLNH ILCEPFMLNH FFCEPYMLNH	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG PDSFDWREHG PDSFDWREHG PESFDWREHG PESFDWREHG PESFDWREKG EPEDKYPYEA ESEDKYPYDA EPEDAYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESEDRYPYDG	AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGN AVTAVKNQGS AVTEVKNQGS AVTEVKNQGN AVTDVKNQGS AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM AVTEVKNQGN B AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN KGECCHIVRS RGETCHLVRS RGETCHLVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KNIPYWIIKN KNKPYWIIKN KNKPYWIIKN		3000 GN I EGQWFLA GN VEGQWFLA GN I EGQWFLA GN VEGQWFLA GN VEGQWFLA GN VEGQWFLA GN VEGQWFLA 1 ELPHDEEKQMQ ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK CLPHDEESMK ELPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESK	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS CRINR-FPTS CGVHEM-ATS CGVHEM-ATS	3200 I 3200 I ALLDCDVVDE ELVDCDKLDQ ELVDCDKLDQ ELVDCDKVDD ELVDCDKVDD ELVDCDKVDQ ELVDCDKLDQ ELVDCDKLDQ ELVDCDKLDQ IGITVDDIQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINATPLQF	213 173 310 310 303 191 252 299 212 293 253 390 386 383 263 2379 286
Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Ac-PFCP Tci-CF-1 Dp-HP-2 Ac-PFCP Sv-UP Aco-UP Dy-CF-1 Aco-UP Sv-UP Aco-UP Dp-HP-2 Ac-PFCP Tci-CF-1 Dp-HP-1 Hc-Pi-1 Aco-UP Dy-HP-2 Aco-UP Dy-HP-1 Aco-UP Dy-HP-1 Hc-Pi-1 Dy-HP-1 Aco-UP Dy-HP-1	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRWVDLT IFPNKIVDLE IYPNKIADLE IFPNKIVDLE IIPNRMANLE IYPNKIADLE GCNGGFPLDA GCEGGLPSQA GCNGGLPSQA	260 I AEGVDPKEPL AEGVYVEDAL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINTTDDY AEGINVTDY AEGINVTDY AEGINA	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG PDSFDWREHG PDSFDWREHG PESFDWREHG PESFDWREHG PESFDWREKG EEDKYPYEA ESEDKYPYDA EPEDAYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA ESEDRYPYDG ESEDRYPYDA ESEDRYPYDG ESEDRYPYDA ESEDRYPYDA ESEDRYPYDG ESEDRYPYDA ESEDRYPYDG ESEDRYPYDA ESEDRYPYDA ESEDRYPYDA ESEDRYPYDA ESEDRYPYDA ESEDRYPYDA ESEDRYPYDA	200 200 1 AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM AVTEVKNQGN 300 1 KAEQCRLVPS RGESCHVAKP KGETCHIVRS RGEECHIVRS RGEECHIVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KGETCHLVRS RGESCHVAKP KNKPFWIIKN KNKPFWIIKN KNKPFWIIKN KNKPFWIIKN		3000 GN I EGQWFLA GN VEGQWFLA GN I EGQWYLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN VEGQWFLA GN VEGQWFLA GN VEGQWFLA GN VEGQWFLA GN VEGQWFLA GN VEGQWFLA S80 L PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK EL PHDEESMK CH PHDESMK CH PH	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKLVSL	3200 GL LDCD VVDE EL VDCDK LDQ EL VDCDK LDQ EL VDCD I IDQ EL VDCD I IDQ EL VDCD I IDQ EL VDCD I VDQ EL VDCD I VDQ EL VDCD I VDQ EL VDCD I VDQ GL VDCD I VDQ EL VDCD I VDQ IGINANPLQF IGINANPLQF IGINANPLQF IGINATPLQF IGI	213 173 310 310 303 191 252 299 212 293 253 390 386 383 263 332 379 286
Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Aco-UP Dp-HP-2 Ac-PFCP Tci-CF-1 Dp-HP-2 Ac-PFCP Tci-CF-1 Hc-PI-2 Dv-CF-1 Dp-HP-2 Ac-PFCP Tci-CF-1 Dp-HP-2 Ac-PFCP Tci-CF-1 Dp-HP-2 Ac-PFCP Sv-UP	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRWVDLT IFPNKIVDLE IYPNKIADLE IFPNKIVDLE IIPNRMANLE IPNKIADLE GCNGGFPLDA GCCGGLPSQA GCNG GCNG GCNG GCNG GCNG GCNG GCNG GCN	260 I AEGVDPKEPL AEGVYVEDAL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINTTDDY AEGINVTDY AEGINV	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG PDSFDWREHG PDSFDWREHG PESFDWREHG PESFDWREKG EPEDKYPYEA ESEDKYPYDA EPEDAYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESEDRYPYDA EPEDAYPYDG GALLVGYGVE GVLIVGYGSE GVLIVGYGSE GVLIVGYGSE	200 1 AVTKVKTEGH AVTSVKNQGM AVTAVKNQGN AVTAVKNQGN AVTOVKNQGS AVTEVKNQGN AVTDVKNQGS AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM AVTEVKNQGN 300 1 KAEQCRLVPS RGESCHVAKP KGETCHIVRS RGEECHIVRS RGEECHIVRS RGETCHIVRS RGETCHIVRS RGETCHIVRS RGESCHVAKP KGETCHIVRS RGESCHVAKP KGETCHIVRS RGESCHVAKP KGETCHIVRS RGESCHVAKP KGETCHIVRS RGESCHVAKP KGETCHIVRS RGESCHVAKP KGETCHIVRS RGESCHVAKP KGETCHIVRS RGESCHVAKP KGETCHIVRS RGESCHVAKP KGETCHIVRS RGESCHVAKP KNKPFWIIKN KN		3000 GN I EGQWFLA GN VEGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN I EGQWFLA GN VEGQWFLA GN VEGQWFLA GN VEGQWFLA GN VEGQWFLA CLPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK ELPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDEESMK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK CLPHDESKK	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KHQLVSLSEQ AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS CRINR-FPTS CGVHEM-ATS CGVHEM-ATS CGVHEM-ATS	3200 GL LDCD VVDE EL VDCDK LDQ EL VDCDK LDQ EL VDCD I IDQ EL VDCD I IDQ EL VDCD I VDQ EL VDCD I VDQ EL VDCD I VDQ EL VDCD I VDQ EL VDCD I VDQ GL VDCD VDQ GL VDCDK LDQ I GI TVDD I QF I GI NANPLQF I	213 173 310 306 303 191 252 299 212 293 253 390 386 383 263 332 379 286
Tci-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Aca-HP Sv-UP Dp-HP-2 Ac-PFCP Tci-CF-1 Hc-PI-1 Hc-PI-2 Dv-CF-1 Dp-HP-2 Aca-HP Sv-UP Aca-HP Sv-UP Dp-HP-2 Aca-FCP Tci-CF-1 Dp-HP-1 Hc-PI-1 Hc-PI-2 Dv-CF-1 Hc-PI-1 Hc-PI-2 Dv-CF-1 Hc-PI-1 Hc-PI-2 Dv-CF-1 Hc-PI-2 Dv-CF-1 CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Dp-HP-1 Hc-PI-2 Dv-CF-1 Dp-HP-2 Aca-PFCP	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IYPNRIADLA IYPNRIADLE IPNKIVDLE IYPNKIADLE IFPNKIVDLE IIPNRMANLE IFPNKIADLE CONGGFPLDA GCNGGFPLDA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA GCNGGLPSQA CONGCLPSQA CONGCLPSQA C	260 I AEGVDPKEPL AEGVYVEDAL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGINTTDDY AEGINVTDY AEGINVTDY	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PDSFDWRDHG PDSFDWREHG PDSFDWREHG PESFDWREKG PESFDWREKG PESFDWREKG EEEDKYPYDA EPEDAYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG ESEDRYPYDA EPEDAYPYDG CALLVGYGE GVLIVGYGSE GVLIVGYGE CVLIVGYGTE	200 1 AVTKVKTEGH AVTSVKNQGM AVTSVKNQGM AVTAVKNQGN AVTAVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGM AVTEVKNQGM AVTEVKNQGM AVTEVKNQGN AVTEVKNQG		3000 GN I EGQWFLA GN I EGQWFLA CH EGQWFLA GN I EGQWFLA S800 1 EGQWFLA GN I EGQWFLA GN I EGQWFLA S800 1 EGQWFLA 1 EGQWFLA S800 1 EGQWFLA 1	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KHQLVSLSEQ AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS AWLVKKGPIS CRINR-FPTS CGVHEM-ATS CGVHEM-ATS CGVHEM-ATS CGVHEM-ATS	3200 I 320 QLLDCDVVDE ELVDCDKLDQ ELVDCDIDQ ELVDCDIDQ ELVDCDIDQ ELVDCDIDQ ELVDCDIDQ ELVDCDIQ ELVDCDIQ IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINATPLQF IGINATPLQF <	213 173 310 306 303 191 252 299 212 293 253 390 386 383 390 386 383 379 286
Тсі-СF-1 Dp-HP-1 Hс-PI-2 Dv-CF-1 Aca-HP Sv-UP Dp-HP-2 Ac-PFCP Tсі-CF-1 Dp-HP-1 Hс-PI-2 Dv-CF-1 Aca-HP Sv-UP Dp-HP-2 Ac-PFCP Tсі-CF-1 Dp-HP-1 Hc-PI-2 Ac-PFCP Tсі-CF-1 Dp-HP-1 Hc-PI-2 Dy-CF-1 Hc-PI-2 Dy-CF-1 Hc-PI-2 Dy-CF-1	DHPNRIVDLA IIPNRVANLE IYPNRIADLA IYPNRIADLA IVPNRMVDLT IFPNKIVDLE IYPNKIADLE IFPNKIVDLE IIPNRMANLE IYPNKIADLE GCNGGFPLDA GCEGGLPSQA GCNGCN GCNG GCNG GCNG GCNG GCNG GCNG G	260 I AEGVDPKEPL AEGVYVEDAL AEGISLDDPL AEGISLDDPL AEGISLDDPL AEGINTTDDY AEGVYVEDAL AEGINVTDDY AEGVYVEDAL AEGIDVKEDL 300 310 311 314 314 316 316 316 317 316 316 316 316 316 316 316 316	PESFDWREHG PETFDWREHG PDSFDWRTHG PDSFDWRTHG PESFDWRDHG PESFDWREHG PESFDWREHG PESFDWREHG PESFDWREKG EPEDKYPYDA ESEDYPYDG ESEDYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG ESESDYPYDG GALLVGYGVE GVLIVGYGSE GVLIVGYGSE GVLIVGYGTE GVLIVGYGTE GVLIVGYGTE	200 1 AVTKVKTEGH AVTSVKNQGM AVTSVKNQGN AVTAVKNQGN AVTAVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN AVTEVKNQGN 1 KAEQCRLVPS RGESCHVAKP KGETCHIVRS RGECCHIVRS RGECCHIVRS RGECCHIVRS RGECCHIVRS RGECCHIVRS RGECCHIVRS RGECCHIVRS RGECCHIVRS RGECCHIVRS RGECCHIVRS RGESCHVAKP KGETCHIVRS RGESCHVAKP KNKPFWIIKN KNKPFWIIKN KNKPFWIIKN RRKPFWIIKN KNKPFWIIKN		3000 GN I EGQWFLA GN VEGQWFLA GN I EGQWFLA GN VEGQWFLA GN V	KKKLVSLSAQ KKKLVSLSEQ KNKLVSLSEQ KNKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKKLVSLSEQ KKLVSL	320 I 320 QLLDCDVVDE ELVDCDKUDQ ELVDCDIDQ ELVDCDIDQ ELVDCDIDQ ELVDCDIDQ ELVDCDIDQ ELVDCDIDQ ELVDCDIQ ELVDCDIQ ELVDCDVDQ IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINANPLQF IGINATPLQF IGINATPLQF <td>213 173 310 306 303 191 252 293 253 390 386 383 3263 332 379 286</td>	213 173 310 306 303 191 252 293 253 390 386 383 3263 332 379 286

Figure 7. Multiple sequence alignment of Tci-CF-1 with the 9 closest homologous sequences. Tci-CF-1: *Teladorsagia circumcincta* secreted cathepsin F (GenBank accession no. ABA01328); Dp-HP-1: *Diploscapter pachys* hypothetical protein WR25_25536 (GenBank accession no. PAV60527); Hc-PI-1: *Haemonchus contortus* proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain-containing protein (GenBank accession no. CDJ88889); Hc-PI-2: *H. contortus* proteinase inhibitor I25 and proteinase inhibitor I29 and peptidase C1A domain-containing protein (GenBank accession no. CDJ92562); Dv-CF-1: *Dictyocaulus viviparus* cathepsin F1 (GenBank accession no. AFM37363); Ac-PFCP: *Ancylostoma ceylanicum* papain family cysteine protease (GenBank accession no. EPB70524); Aca-HP: *Angiostrongylus cantonensis* hypothetical protein Angca_010213 (GenBank accession no. KAE9418773); Sv-UP: *Strongylus vulgaris* unnamed protein product (GenBank accession no. VDM81154); Aco-UP: *Angiostrongylus costaricensis* unnamed protein product (GenBank accession no. VDM61191); Dp-HP-2: *D. pachys* hypothetical protein WR25_24125 (GenBank accession no. PAV67875). Gaps indicated by a dash; ERFNAQ, E/DxGTA, GxNxFxD and GCNGG motif residues indicated by square, cross, circle and diamond, respectively; catalytic triad residues indicated by a downward arrow.

more parents correspond to reticulate events such as hybridization or recombination. In this study, the reticulate network in Figure 4B shows that when Tci-CF-1 is selected as the outgroup, the homologous proteins are quite evolutionarily distant because many parent nodes correspond to many possible reticulate events between them. The long edge lengths are indicative of the weight of the associated split and are analogous to the length of a branch in a phylogenetic tree.³⁶ The distance



Figure 8. Homology model of *Teladorsagia circumcincta* secreted cathepsin F (GenBank accession no. ABA01328). (A) Pro-region (line residues), mature domain (ribbon), locations of polymorphisms in variants 1, 2, and 3 (bold side-chain residues), and the catalytic triad (bold side-chain residues) which is exposed following cleavage of the pro-peptide. (B) Magnified view of the active site indicating bonds between the pro-region and catalytic triad residues.



Figure 9. Structural comparison of human and *Teladorsagia circumcincta* cathepsin proteins. X-ray crystallography of mature domains of human cathepsins L⁴⁰ and F,⁴¹ and homology modeling of *T. circumcincta* cathepsin F (GenBank accession no. ABA01328) and the pro-regions of human cathepsins L (UniProtKB accession no. P07711) and F (UniProtKB accession no. Q9UBX1). (A) Human cathepsin F; (B) *T. circumcincta* secreted cathepsin F; (C) Human cathepsin L; pro-region, mature domain and cystatin-like domain highlighted in different colors.

between Tci-CF-1 and its closest homologs is relatively large compared to the distances between the homologs themselves and this complements the %ID between species in Table 1. The increased number of parent nodes illustrates there are many hypothetical ancestral species between these proteins, highlighting how divergent they are.

In summary, *T. circumcincta* cathepsin F has no close homologs even in closely related species such as *H. contortus*

which is a member of the same subfamily.⁵³ The absence of close homologs indicates Tci-CF-1 may have evolved relatively recently, whether because of host immune pressure or other factors leading to rapid change. Cathepsin F has characteristics of both cathepsins F and L. The Tci-CF-1 pro-region contains motifs characteristic of both cathepsins F and L; however, homology modeling indicates that it lacks a cystatin-like domain, making it structurally more similar to cathepsin L.

The bioinformatic investigation of Tci-CF-1 provides insights into the presence of amino acid changes/substitutions, and how these may influence the function of cathepsin F. This information can be used to design powerful functional studies to improve our understanding of the role of cathepsin F in the immunogenicity of *T. circumcincta*.

Acknowledgements

We would like to thank the La Trobe University High Performance Computing cluster team for access to their systems and support. Molecular graphics and analyses performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from the National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases.

Authors' Contributions

SS, CC, and MS conceived of the presented idea. SS developed the theory and performed the computations, research, and wrote the manuscript. CJ and MS verified the analytical methods and supervised the findings of this work. All authors discussed the results and contributed to the final manuscript.

ORCID iD

Sarah Sloan (D) https://orcid.org/0000-0002-2131-4899

Supplemental material

Supplemental material for this article is available online.

REFERENCES

- Blaxter M, Koutsovoulos G. The evolution of parasitism in Nematoda. *Parasitol*ogy. 2015;142(Suppl 1):S26-S39.
- 2. Poinar G. Nematoda (Roundworms). In: eLS. 2012.
- Savage RJG, Long MR. Mammal Evolution: An Illustrated Guide. New York, NY: Facts on File Publications; 1986.
- Bartley DJ, Jackson E, Johnston K, et al. A survey of anthelmintic resistant nematode parasites in Scottish sheep flocks. *Vet Parasitol*. 2003;117:61-71.
- Marchiondo AA, Cruthers LR, Reinemeyer CR. Nematoda. In: Marchiondo AA, Cruthers LR, Fourie JJ, eds. *Parasiticide Screening*, *Volume 2*. London, UK: Academic Press; 2019:135-335.
- Stear MJ, Bishop SC, Henderson NG, Scott I. A key mechanism of pathogenesis in sheep infected with the nematode *Teladorsagia circumcincta*. *Anim Health Res Rev.* 2003;4:45-52.
- Smith WD, Jackson F, Jackson E, et al. Transfer of immunity to Ostertagia circumcincta and IgA memory between identical sheep by lymphocytes collected from gastric lymph. Res Vet Sci. 1986;41:300-306.
- Smith WD, Jackson F, Jackson E, Williams J. Local immunity and Ostertagia circumcincta: changes in the gastric lymph of immune sheep after a challenge infection. J Comp Pathol. 1983;93:479-488.
- Smith WD, Jackson F, Jackson E, Williams J. Age immunity to Ostertagia circumcincta: comparison of the local immune responses of 4 1/2- and 10-month-old lambs. J Comp Pathol. 1985;95:235-245.
- 10. Stear MJ, Bairden K, Bishop SC, et al. The genetic basis of resistance to *Osterta-gia circumcincta* in lambs. *Vet J.* 1997;154:111-119.
- Stear MJ, Doligalska M, Donskow-Schmelter K. Alternatives to anthelmintics for the control of nematodes in livestock. *Parasitology*. 2007;134(Pt 2):139-151.
- Jackson F, Coop RL. The development of anthelmintic resistance in sheep nematodes. *Parasitology*. 2000;120 Suppl:S95-107.
- 13. Nisbet AJ, McNeilly TN, Wildblood LA, et al. Successful immunization against a parasitic nematode by vaccination with recombinant proteins. *Vaccine*.

2013;31:4017-4023.

- Stear MJ, Bishop SC, Doligalska M, et al. Regulation of egg production, worm burden, worm length and worm fecundity by host responses in sheep infected with Ostertagia circumcincta. Parasite Immunol. 1995;17:643-652.
- Chung YB, Kong Y, Joo IJ, Cho SY, Kang SY. Excystment of *Paragonimus westermani* metacercariae by endogenous cysteine protease. *J Parasitol.* 1995;81: 137-142.
- Hashmi S, Britton C, Liu J, Guiliano DB, Oksov Y, Lustigman S. Cathepsin L is essential for embryogenesis and development of *Caenorhabditis elegans. J Biol Chem.* 2002;277:3477-3486.
- Lustigman S, McKerrow JH, Shah K, et al. Cloning of a cysteine protease required for the molting of Onchocerca volvulus third stage larvae. J Biol Chem. 1996;271:30181-30189.
- Carmona C, Dowd AJ, Smith AM, Dalton JP. Cathepsin L proteinase secreted by *Fasciola hepatica* in vitro prevents antibody-mediated cosinophil attachment to newly excysted juveniles. *Mol Biochem Parasitol*. 1993;62:9-17.
- Kang TH, Yun DH, Lee EH, et al. A cathepsin F of adult *Clonorchis sinensis* and its phylogenetic conservation in trematodes. *Parasitology*. 2004;128(Pt 2): 195-207.
- Illy C, Quraishi O, Wang J, Purisima E, Vernet T, Mort JS. Role of the occluding loop in cathepsin B activity. J Biol Chem. 1997;272:1197-1202.
- Turk V, Stoka V, Vasiljeva O, et al. Cysteine cathepsins: from structure, function and regulation to new frontiers. *Biochim Biophys Acta*. 2012;1824:68-88.
- Nagler DK, Sulea T, Menard R. Full-length cDNA of human cathepsin F predicts the presence of a cystatin domain at the N-terminus of the cysteine protease zymogen. *Biochem Biophys Res Commun.* 1999;257:313-318.
- Redmond DL, Smith SK, Halliday A, et al. An immunogenic cathepsin F secreted by the parasitic stages of *Teladorsagia circumcincta*. Int J Parasitol. 2006;36:277-286.
- Vernet T, Berti PJ, de Montigny C, et al. Processing of the papain precursor. The ionization state of a conserved amino acid motif within the pro region participates in the regulation of intramolecular processing. *J Biol Chem.* 1995;270: 10838-10846.
- Collins PR, Stack CM, O'Neill SM, et al. Cathepsin L1, the major protease involved in liver fluke (*Fasciola hepatica*) virulence: propetide cleavage sites and autoactivation of the zymogen secreted from gastrodermal cells. *J Biol Chem.* 2004;279:17038-17046.
- Smith AM, Dowd AJ, Heffernan M, Robertson CD, Dalton JP. Fasciola hepatica: a secreted cathepsin L-like proteinase cleaves host immunoglobulin. Int J Parasitol. 1993;23:977-983.
- Nisbet AJ, Redmond DL, Matthews JB, et al. Stage-specific gene expression in *Teladorsagia circumcincta* (Nematoda: Strongylida) infective larvae and early parasitic stages. *Int J Parasitol.* 2008;38:829-838.
- Howe KL, Bolt BJ, Shafie M, Kersey P, Berriman M. WormBase ParaSite a comprehensive resource for helminth genomics. *Mol Biochem Parasitol*. 2017;215:2-10.
- Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database C. The sequence read archive. *Nucleic Acids Res.* 2011;39(Database issue):D19-D21.
- Vernet T, Tessier DC, Chatellier J, et al. Structural and functional roles of asparagine 175 in the cysteine protease papain. *J Biol Chem.* 1995;270:16645-16652.
- Blom N, Sicheritz-Pontén T, Gupta R, et al. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*. 2004;4(6):1633-1649.
- Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 2006;23:254-267.
- Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4:406-425.
- Huson DH, Rupp R, Scornavacca C. Phylogenetic Networks: Concepts, Algorithms and Applications. Cambridge: Cambridge University Press; 2010.
- Hamming RW. Error detecting and error correcting codes. Bell Syst Tech J. 1950;29:147-160.
- Bryant D, Moulton V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol*. 2004;21:255-265.
- Dress AW, Huson DH. Constructing splits graphs. IEEE/ACM Trans Comput Biol Bioinform. 2004;1:109-115.
- Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc.* 2015; 10:845.
- Goddard TD, Huang CC, Meng EC, et al. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* 2018;27:14-25.
- Hardegger LA, Kuhn B, Spinnler B, et al. Halogen bonding at the active sites of human cathepsin L and MEK1 kinase: efficient interactions in different environments. *ChemMedChem*. 2011;6:2048-2054.
- Somoza JR, Palmer JT, Ho JD. The crystal structure of human cathepsin F and its implications for the development of novel immunomodulators. *J Mol Biol.* 2002;322:559-568.

- Barrett AJ, Rawlings ND. Evolutionary lines of cysteine peptidases. *Biol Chem.* 2001;382:727-733.
- Deussing J, Tisljar K, Papazoglou A, Peters C. Mouse cathepsin F: cDNA cloning, genomic organization and chromosomal assignment of the gene. *Gene*. 2000;251:165-173.
- Groves MR, Taylor MAJ, Scott M, Cummings NJ, Pickersgill RW, Jenkins JA. The prosequence of procaricain forms an α-helical domain that prevents access to the substrate-binding cleft. *Structure*. 1996;4:1193-1203.
- Stear MJ, Singleton D, Matthews L. An evolutionary perspective on gastrointestinal nematodes of sheep. J Helminthol. 2011;85:113-120.
- Stear MJ, Strain S, Bishop SC. Mechanisms underlying resistance to nematode infection. *Int J Parasitol.* 1999;29:51-56; discussion 73-55.
- Chilton NB, Newton LA, Beveridge I, Gasser RB. Evolutionary relationships of trichostrongyloid nematodes (Strongylida) inferred from ribosomal DNA sequence data. *Mol Phylogenet Evol.* 2001;19:367-386.
- Chilton NB, Huby-Chilton F, Gasser RB, Beveridge I. The evolutionary origins of nematodes within the order Strongylida are related to predilection sites within hosts. *Mol Phylogenet Evol.* 2006;40:118-128.
- Chilton NB, Huby-Chilton F, Koehler AV, Gasser RB, Beveridge I. The phylogenetic relationships of endemic Australasian trichostrongylin families (Nematoda: Strongylida) parasitic in marsupials and monotremes. *Parasitol Res.* 2015;114:3665-3673.
- Kolkman JA, Stemmer WPC. Directed evolution of proteins by exon shuffling. Nat Biotechnol. 2001;19:423-428.
- Betts MJ, Russell RB. Amino Acid Properties and Consequences of Substitutions. In: Barnes MR, Gray IC, eds. *Bioinformatics for Geneticists*. Chichester, UK: John Wily & Sons, Ltd.; 2003:289-316.
- Karrer KM, Peiffer SL, DiTomas ME. Two distinct gene subfamilies within the family of cysteine protease genes. *Proc Natl Acad Sci USA*. 1993;90:3063-3067.
- Parkinson J, Mitreva M, Whitton C, et al. A transcriptomic analysis of the phylum Nematoda. *Nat Genet*. 2004;36:1259-1267.