

**Multidisciplinary Investigation of Cathepsin F and
Implications for Improving Control of *Teladorsagia
circumcincta***

Submitted by Sarah Sloan

Bachelor of Science; Bachelor of Animal and Veterinary Biosciences with Honours

A thesis submitted in total fulfilment of the requirements for the degree of

Doctor of Philosophy; School of Animal, Plant and Soil Sciences; College of Science, Health and Engineering

La Trobe University, Victoria, Australia

November 2021

Contents

Contents.....	i
List of Figures.....	iv
List of Tables.....	iv
List of Supplementary Files.....	v
Abstract.....	vi
Statement of Authorship.....	vii
Scholarship Acknowledgement.....	viii
Acknowledgements.....	ix
Thesis Preface.....	x
Chapter 1.....	1
Introduction.....	1
1.1 <i>Teladorsagia circumcincta</i>	2
1.1.1 Physiology.....	2
1.1.2 Lifecycle.....	3
1.1.3 Impact.....	4
1.1.4 Infection.....	5
1.1.5 Control methods.....	6
1.1.5.1 Anthelmintics.....	6
1.1.5.2 Pasture management.....	6
1.1.5.3 Biological controls.....	6
1.1.5.4 Nutritional supplementation.....	6
1.1.5.5 Vaccines.....	7
1.1.5.6 Breeding.....	7
1.2 Cysteine Proteases.....	8
1.2.1 Cathepsin L.....	9
1.2.2 Cathepsin B.....	9
1.2.3 Cathepsin F.....	9
1.3 DNA extraction.....	10
1.3.1 Chelating.....	11
1.3.2 Precipitation.....	11
1.3.3 Silica binding.....	11
1.3.4 DNA analysis.....	12
1.4 Genome sequencing.....	12
1.4.1 Sequencing methods.....	13
1.4.1.1 Illumina.....	13

1.4.1.2 PacBio	13
1.4.1.3 Oxford Nanopore	14
1.4.2 <i>Teladorsagia circumcincta</i> genome	15
1.5 Research Plan	16
Chapter 2	17
Cathepsin F of <i>Teladorsagia circumcincta</i> is a recently evolved cysteine protease	17
2.2 Publication details	19
2.3 Statement of contribution of joint authorship	19
2.4 Statement from the co-author confirming the authorship contribution of the PhD candidate	20
2.5 Manuscript.....	21
Chapter 3	33
Comparative evaluation of different molecular methods for DNA extraction from individual <i>Teladorsagia circumcincta</i> nematodes	33
3.2 Publication details	35
3.3 Statement of contribution of joint authorship	35
3.4 Statement from the co-author confirming the authorship contribution of the PhD candidate	36
3.5 Manuscript.....	37
Chapter 4	50
<i>Teladorsagia circumcincta</i> genome sequencing on the MinION™ and microbiome species identification	50
4.1 Introduction	51
4.2 Materials & Methods	52
4.2.1 Sampling	52
4.2.2 DNA extraction	52
4.2.3 Sequencing on MinION™.....	53
4.2.4 Read classification	53
4.2.5 Genome assembly	53
4.2.6 Cathepsin F gene	53
4.3 Results.....	54
4.3.1 Sequence basecalling analysis	54
4.3.2 WIMP read classification.....	55
4.3.3 Microbiome analysis of TcM1-3 classified reads.....	56
4.3.4 Genome assembly	57
4.3.5 Cathepsin F gene	60
4.4 Discussion	67
4.4.1 Genome assembly	67

4.4.2 Microbiome analysis	68
4.5 Conclusion	72
4.6 Supplementary Information	73
Chapter 5	74
Preliminary cloning and expression of the secreted cathepsin F protein from <i>Teladorsagia circumcincta</i> larvae.....	74
5.1 Introduction.....	75
5.2 Materials & Methods	76
5.2.1 Cloning and transformation of TcCatF cDNA in pPICZα B.....	76
5.2.2 Preparation of recombinant proteins from yeast	77
5.2.3 Confirmation of recombinant protein expression	79
5.2.4 Protein activation analysis.....	80
5.3 Results.....	81
5.3.1 Characterisation of the full-length prev-TcCatF cDNA clone and predicted protein	81
5.3.2 Expression of prev-TcCatF in <i>Pichia pastoris</i>	83
5.3.3 Activation of prev-TcCatF.....	85
5.4 Discussion	85
5.5 Conclusion	88
Chapter 6	89
General Discussion	89
6.1 Introduction.....	90
6.2 Cathepsin F bioinformatic analysis	90
6.3 Individual nematode DNA extraction	91
6.4 MinION™ sequencing of <i>T. circumcincta</i>	92
6.5 Cathepsin F gene	93
6.6 <i>T. circumcincta</i> microbiome analysis	94
6.7 Cathepsin F protein expression	95
6.8 Conclusion	96
References	97
Appendix.....	115
S2.1 Chapter 2 supplementary data 1	115
S2.2 Chapter 2 supplementary data 2	122
S2.3 Chapter 2 supplementary data 3	123
S3.1 Chapter 3 supplementary file 1.....	132
S3.2 Chapter 3 supplementary file 2.....	133

List of Figures

Figure 1.1: Light microscopy of a cluster of <i>Teladorsagia circumcincta</i>	3
Figure 1.2: Standard one-host lifecycle of <i>Teladorsagia circumcincta</i>	4
Figure 4.1: Bar graph of the number of reads for each species present in “What’s In My Pot?” <i>Homo sapiens</i> classified reads which were BLASTn against the NCBI nucleotide database	56
Figure 4.2: Bar graph indicating the length (kbp) of 9,129 sequences in the <i>Teladorsagia circumcincta</i> genome assembly in this study, and the number of occurrences for that size.....	60
Figure 4.3: Contig00000245 with Augustus predicted genes (g1092 – 1112) and annotated to include <i>Teladorsagia circumcincta</i> cathepsin F gene fragments and gene	62
Figure 4.4: Contig00000282 with Augustus predicted genes (g1363 – 1371) and annotated to include <i>Teladorsagia circumcincta</i> cathepsin F gene fragments and gene	62
Figure 4.5: Multiple sequence alignment of <i>Teladorsagia circumcincta</i> secreted cathepsin F, and the cathepsin F genes identified in this study	63
Figure 5.1: Map of the expression vector pPICZα B - TcCatF.....	78
Figure 5.2: Codon-optimised, commercially synthesised TcCatF cDNA sequence used in the pPICZα B construct.....	82
Figure 5.3: TcCatF and α-factor secretion signal amino acid sequence translated from pPICZα B + TcCatF construct	83
Figure 5.4: SDS-PAGE analysis of the expression culture supernatant using method 2	84
Figure 5.5: SDS-PAGE analysis of prev-TcCatF 2 purified protein at varying pH and incubation times	85

List of Tables

Table 4.1: Oxford Nanopore MinION™ sequencing data and What’s In My Pot? read classification data.	54
Table 4.2: Presence of phyla and classes present in TcM1-3 classified sequence reads	57
Table 4.3: Results of BLASTn of Augustus predicted genes on contigs 245 and 282.	64
Table 5.1: Starter and expression culture setup methods for production of prev-TcCatF	79

List of Supplementary Files

S4.1 Supplementary File 1: Genome assembly contigs of the <i>Teladorsagia circumcincta</i> canu assembly created using Oxford Nanopore Technology reads sequenced in this study and PacBio RS reads retrieved from the Sequence Read Archive	73
S4.2 Supplementary File 2: General feature format 3 file of the Augustus gene predictions for the <i>Teladorsagia circumcincta</i> assembly created in this study	73
S2.1 Chapter 2 Supplementary File 1: Alignment of the mRNA sequence for <i>Teladorsagia circumcincta</i> cathepsin F against the <i>T. circumcincta</i> genome in WormBase ParaSite.....	115
S2.2 Chapter 2 Supplementary File 2: Alignment of TELCIR_06733 and _06734 translated exons from draft <i>Teladorsagia circumcincta</i> genome in WormBase Parasite against <i>T. circumcincta</i> secreted cathepsin F	122
S2.3 Chapter 2 Supplementary File 3: Phyre ² secondary structure and disorder prediction, detailed template information, and domain analysis for <i>Teladorsagia circumcincta</i> cathepsin F	123
S3.1 Chapter 3 Supplementary File 1: Table of statistically significant differences in NanoDrop 2000™ and Qubit™ DNA concentration, and 260/230 nm absorbance ratio using the 11 DNA extraction methods according to the Dunn's Multiple Comparison Test.....	132
S3.2 Chapter 3 Supplementary File 2: DNA extraction methods.....	133

Abstract

Teladorsagia circumcincta is the most important parasitic nematode of sheep in temperate regions worldwide due to its economic effect on production and animal welfare. Cathepsin F is the most abundant protein secreted by fourth-stage larvae, a critical time for nematode establishment within the host, however, little is known about its role or function.

A multidisciplinary approach was utilised to study cathepsin F of *T. circumcincta*. The cathepsin F protein sequence, tertiary structure, and phylogeny were explored using bioinformatic analyses. Pooled nematode DNA samples have been critical for *T. circumcincta* genome construction, but individual specimens are required for determination of genetic variation within populations. A method for DNA extraction from individual *T. circumcincta* specimens was determined. The resulting DNA was sequenced using Oxford Nanopore Technology (ONT) for the first time, and the reads were assembled into a draft genome. The cathepsin F gene was further analysed and annotated. Additional non-*T. circumcincta* genetic sequences filtered out of the genome assembly were classified and analysed at their Phylum and Class level using ONT's What's In My Pot? pipeline, providing insight into the microbiome of sheep infected with *T. circumcincta* and the microbiome of *T. circumcincta* itself. Production of a recombinant cathepsin F protein for functional and structural studies was attempted. However, difficulties due to Covid-19 were experienced, producing an active protein was not achieved, and this chapter was ultimately abandoned.

This research has enhanced the ability of researchers to study the control of *T. circumcincta* by providing background information on the cathepsin F protein to help with future functional and structural research, providing a proven method for extracting DNA from individual specimens, and improving the genomic information available through additional sequencing data and microbiome analysis.

Statement of Authorship

This thesis consists primarily of work by the author that has been published or submitted for publication as described in the text. Except where reference is made in the text of the thesis, this thesis contains no other material published elsewhere or extracted in whole or in part from a thesis submitted for the award of any other degree or diploma. No other person's work has been used without due acknowledgement in the main text of the thesis. This thesis has not been submitted for the award of any degree or diploma in any other tertiary institution.

Sheep work was approved by the Animal Ethics Committee of Federation University (17-008).

Sarah Sloan

1st November 2021

Scholarship Acknowledgement

This work was supported by an Australian Government Research Training Program Scholarship and also a La Trobe University Postgraduate Research Scholarship for Sarah Sloan.

Acknowledgements

I have had some fantastic experiences during this PhD journey, and it would not have been possible without those who supported me over the last three-ish years.

My gratitude to my supervisor Prof. Michael Stear for his continual support, encouragement, and patience throughout my PhD. Thank you for selecting me for this project, believing in my capabilities to complete this PhD, and letting me steer my course while always providing the much-needed suggestions and advice when needed.

Dr. Caitlin Jenvey, I want to thank you for all the support that you have provided me over the last few years. It was invaluable.

Prof. Travis Beddoe, thank you for your constant support and advice over the course of my PhD.

A special thank you to the Stear and Beddoe lab groups, who have taught me a lot throughout my studies and whose support I deeply appreciate. Specifically, Gemma Zerna, Dr. Tim Cameron, James O'Sullivan, Lily Tran, and Nur Nasuha. To everyone else who assisted me through my thesis, thank you. Also to Tongda Li, Tim Sawbridge, and Reannon Smith for their help throughout.

I want to thank my family for their continual support, patience, and encouragement over the last nine years throughout my undergraduate, honours and PhD studies. It has been a long slog and I appreciate all the assistance and support that you have provided me at different times over the years to help me out when I needed it. Particularly I want to thank my Mum, Jacky Edwards, for going above and beyond, and being an unofficial supervisor.

To my friends Rose Albiston, Caroline Bell, Kevin Chen, Dylan and Tomecka Ellis, Maxime Hexter, Marty Hyde, Daniel James, Amrith Krishnaswamy, Etai Krispin, Aaron Lombardo, Yasmine Luu, Daniel Orchard, Trisha Reibelt, Andrew Spreadbury, Sophie Westland and Belinda Wigg, you have all been an incredible support base since we first met, and your continued encouragement has been invaluable over these last few years and months.

I am very thankful to my partner, Scott Walker, for his endless support, love, encouragement, computer technical ability, coding expertise, and patience with me throughout my PhD journey.

And finally, Chloe and Nova. I want to thank my furry family for your companionship over the last 2 years. You gave me a reason to take breaks and take walks that have cleared my thoughts. Your unconditional love and your stress relief have kept me going while working from home and completing this thesis.

Thesis Preface

This thesis consists of 6 chapters, with the original experimental research presented in the form of two peer-reviewed journal articles and two manuscripts that are presented in the thesis submission format. Chapter 1 provides a general overview of the literature in this area of research. The two published manuscripts are presented in Chapter 2 and Chapter 3 and with the final experimental manuscripts presented in Chapter 4 and Chapter 5.

The research chapters are structured with their own introduction, methodology, results, and discussion sections. Also, the research chapters are prefaced by a summary of the research completed, the manuscript publication details, contribution of the co-authors and a statement from the co-author confirming the authorship contribution of the PhD candidate. Chapter 6 provides a general discussion that integrates the major themes from the three manuscripts as well as providing suggestions for future research.

The 2 published experimental chapters are presented with the respective referencing, citation and formatting styles of the corresponding journals. A single referencing and citation style (APA 6th) has been employed for chapters 1, 4, 5, and 6, and the reference list is provided at the end of this thesis.

Chapter 1

Introduction

1.1 *Teladorsagia circumcincta*

Nematoda are such an ancient group that most terrestrial plants and larger animals are associated with at least one parasitic nematode (Blaxter *et al.*, 2015). *Teladorsagia circumcincta* is a nematode from the Order Rhabditida and Family Trichostrongylidae, and more specifically belongs to Strongyloidea of clade V, alongside *Ostertagia ostertagi* and *Haemonchus contortus* (Parkinson *et al.*, 2004). *T. circumcincta* is the most important parasitic nematode of sheep and goats in temperate areas worldwide (Bartley *et al.*, 2003), was previously known as *Ostertagia circumcincta*, and is colloquially known as the Brown Stomach Worm (Taylor *et al.*, 2007).

Fossilisation of various nematode species occurred as early as 135 million years ago (mya; Poinar, 2012), however, there are no records of fossilised *T. circumcincta*, and as such an estimate of their time of divergence is difficult to calculate. The earliest bovids appeared 20 mya (Janis, 1993; Savage *et al.*, 1986) and during this time the climate was warming and becoming drier, leading to the development of open woodland, grasslands (Janis, 1993) and grazing Bovidae (Stebbins, 1981). *T. circumcincta* is observed in sheep and goats but not cattle, however, the closely related nematode *O. ostertagi* is observed in cattle, but not sheep and goats. We can reasonably assume the divergence of these nematode species occurred within this 20 million year time-frame, likely alongside their hosts (Stear *et al.*, 2011).

1.1.1 Physiology

T. circumcincta are slender, reddish-brown worms with a short buccal cavity (Taylor *et al.*, 2007). Their size varies considerably amongst sheep, as worm length is affected by host immune pressure (Stear *et al.*, 1999a), but typically female size ranges from 8 – 10 mm, and males 6 – 8 mm long (Taylor *et al.*, 2007). The basic body plan of *T. circumcincta* can be described as an outer and inner tube (Figure 1.1). A tough, flexible, collagenous cuticle covers the epidermal cells of the outer tube. The inner tube is the digestive tract; a muscular pharynx and an intestine which runs from the pharynx to the anus. Between the tubes is the pseudocoelom, which in adults is occupied by gonads. Additionally, males have a copulatory bursa; a modification of the posterior end used for breeding (Wood, 2002b).



Figure 1.1: Light microscopy of a cluster of *Teladorsagia circumcincta*. A: head; B: digestive tract; C: copulatory bursa of a male; D: egg within a female. Photograph by S. Sloan.

1.1.2 Lifecycle

T. circumcincta has a standard one-host lifecycle (Figure 1.2). Eggs are passed in faeces onto pasture where they develop into pre-infective larval stages (L1 and L2). L1 and L2 feed on microorganisms in the soil and develop into infective third-stage larvae (L3), which are ingested by the host. Survival and development of these external larval stages is largely determined by the humidity and temperature of faeces and soil (Pandey *et al.*, 1993). L3 exsheath in the rumen and enter the lumen of the abomasum glands 2-3 days after ingestion, where they develop into the pre-adult (L4) and immature adult (L5) stages, and establish within the host. L5 then mature into sexually active adults approximately 12 or more days after infection, which feed and breed on the mucosal surface of the abomasa (Marchiondo *et al.*, 2019; Venturina *et al.*, 2013). In a naïve host, ingested larvae take as little as 14 days to develop into egg-producing adults. Resistant hosts are able to delay maturation of larvae for at least 8 weeks (Stear *et al.*, 1995b).

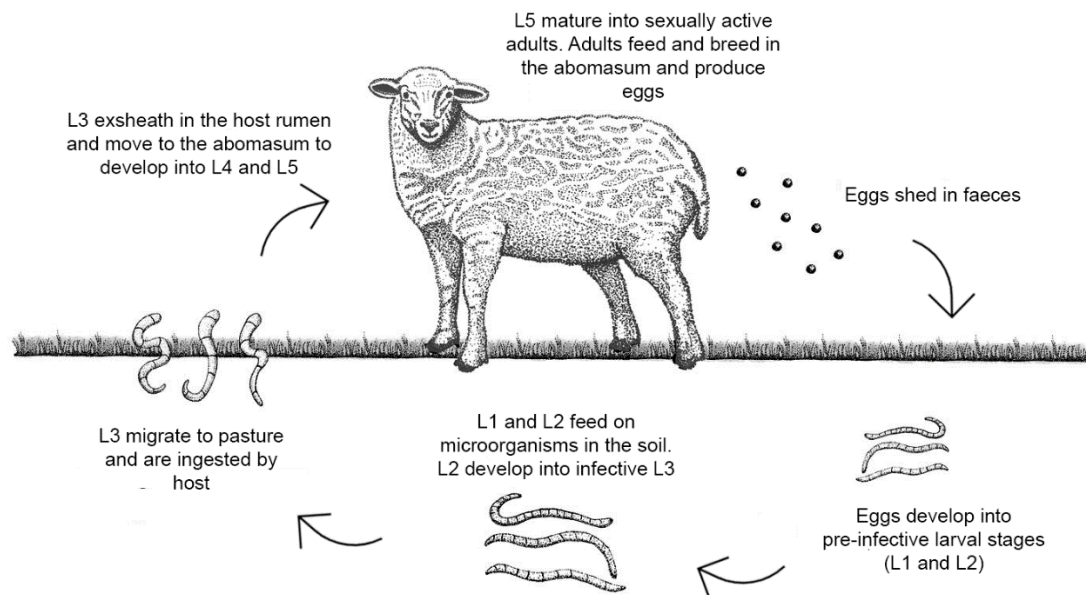


Figure 1.2: Standard one-host lifecycle of *Teladorsagia circumcincta*. Figure adapted from *Farm Health First* (2021).

Parasitic infections of livestock are of major socio-economic importance worldwide. Estimates demonstrate internal parasites of sheep alone cost \$436 million annually in Australia (Lane *et al.*, 2015). *T. circumcincta* is seen worldwide, but is most prevalent in cool, temperate areas, such as south-eastern and south-western Australia and the United Kingdom (Bartley *et al.*, 2003; Burgess *et al.*, 2012; Grillo *et al.*, 2007; Leathwick *et al.*, 2014; O'Connor *et al.*, 2006; Stear *et al.*, 1998). *T. circumcincta* has also been confirmed in Argentina (Eddi *et al.*, 1996), Brazil (Echevarria *et al.*, 1996), Canada (Hoberg *et al.*, 1999), Canary Islands (Gonzalez *et al.*, 2019), Egypt (Elseadawy *et al.*, 2019), Europe (Papadopoulos *et al.*, 2012), France (Gruner *et al.*, 2004), Greece (Papadopoulos *et al.*, 2001), Iran (Ashrafi *et al.*, 2020), Ireland (Keegan *et al.*, 2017), Italy (Traversa *et al.*, 2007), The Netherlands (Borgsteede *et al.*, 2010), New Zealand (Choi *et al.*, 2017; Palevich *et al.*, 2019), Pakistan (Muhammad *et al.*, 2015), Slovak Republic (Cernanska *et al.*, 2006), Spain (Martinez-Valladares *et al.*, 2012a; Martinez-Valladares *et al.*, 2012b), and Uruguay (Nari *et al.*, 1996), among others.

The costs associated with *T. circumcincta* infection accumulate due to low product yield (e.g. fleece, meat, and milk), testing, and treatment. However, regular testing of animals and of anthelmintic efficacy on a flock has been shown to be cost-effective even when infection is subclinical because parasite resistance makes anthelmintics less effective (Miller *et al.*, 2012).

1.1.4 Infection

Infection with *T. circumcincta* leads to Teladorsagiosis, and can result in production losses, reduced animal welfare, parasitic gastroenteritis, poor growth performance, and weight loss (Stear *et al.*, 2003). Death can result if animals are left untreated, but the major impact of this disease is the reduced growth rate of lambs.

Efficient and welfare-friendly sheep production requires nematode control (Stear *et al.*, 2011). Pathogenesis is largely due to a relative protein deficiency caused by infected animals reduced eating, less efficient digestion of protein, loss of host proteins into the gastrointestinal tract due to breaches in the epithelial barrier, and increased protein demand for immune and inflammatory responses (Stear *et al.*, 2003). Supplementary feeding with protein or urea before and during infection can reduce or prevent clinical signs (Coop *et al.*, 1995; Stear *et al.*, 2000). If infection is severe enough, typical features include increased mucus production, hyperplasia, decreased acid production in the gut, and inappetence. The severity of infection depends on concurrent infections, nutritional status, and immune response genetics (Stear *et al.*, 1995a; Stear *et al.*, 2003; Stear *et al.*, 1999b).

Susceptibility to infection among lambs and kids is variable and much of the variation is due to genetics (Murphy *et al.*, 2010; Stear *et al.*, 1997; Stear *et al.*, 1995a; Stear *et al.*, 1999b). Very young animals are protected by maternal antibodies from suckling of colostrum, and eating very little contaminated grass (Stear *et al.*, 2003). As lambs mature and increase the grass component of their diet, if the grass is contaminated with infective larvae, they begin to develop their own type I hypersensitivity immune response (Stear *et al.*, 1997; Stear *et al.*, 1995a). A major part of ongoing immunity appears to be the IgA-mediated inhibition of worm growth and fertility (fecundity).

IgA plays a crucial role in protection against *T. circumcincta* (Stear *et al.*, 1997). Variation in worm length and fecundity has been shown to be significantly associated with the activity of IgA against L4 in the abomasal mucus. IgG activity, the number of mast cells and the number of globule leucocytes in the abomasal mucosa do not affect worm length and fecundity (Stear *et al.*, 1995a). IgA appears to be the most important regulator of worm fecundity, although some animals with moderately strong anti-parasite IgA responses still have fertile worms (W. D. Smith *et al.*, 1985; Stear *et al.*, 1995a). The ability of some sheep to regulate worm fecundity is possibly due to strong IgA responses to specific parasite molecules which are not recognized by all sheep, and requires further research (McCrie *et al.*, 1997).

1.1.5 Control methods

Several methods are used to control *T. circumcincta* infection including anthelmintic treatment, nutritional supplementation, vaccination, selective breeding, and pasture management.

1.1.5.1 Anthelmintics

Anthelmintic treatment has traditionally been the favoured control method. Several classes of drugs are used to control nematodes in livestock: benzimidazoles, nicotinic agonists, macrocyclic lactones, amino-acetonitrile derivatives, spiroindoles, cyclooctadepsipeptides, and tribendimidine (Abongwa *et al.*, 2017). However, *T. circumcincta* is developing drug resistance (Bartley *et al.*, 2003; Jackson *et al.*, 2000). The three most common anthelmintics are albendazole, ivermectin and levamisole. These drugs are used alone or in tandem, and when used with 'best practise parasite management' have shown nematode reversion towards drug susceptibility (Leathwick *et al.*, 2015).

Albendazole, a benzimidazole, prevents newly hatched larvae from growing or multiplying by inhibiting the assembly of microtubules, preventing absorption of sugar, and preventing the formation of spindle fibres needed for cell division (Capece *et al.*, 2009; Riviere *et al.*, 2009). Ivermectin, a positive allosteric modulator, stimulates excessive release of neurotransmitters in the nervous system of parasites, paralysing the worm or inactivating the parasite gut, leading to death (Martin *et al.*, 2021). Levamisole, a nicotinic agonist, works similarly, with continued stimulation of the parasitic worm muscles leading to paralysis and death (Renoux, 1980).

1.1.5.2 Pasture management

Pasture management includes reducing stocking density to reduce pasture contamination, alternating pasture use between stock and crops, rotating between young and older animals, moving vulnerable animals onto less contaminated fields, and implementing rotational grazing to deprive infective larvae of a host (Stear *et al.*, 2007).

1.1.5.3 Biological controls

Biological control includes using predatory, nematode-trapping fungi such as *Duddingtonia flagrans* to reduce infective larvae numbers (Waller *et al.*, 2004a), as well as some plant species such as chicory (*Cichorium intybus*) and birdsfoot trefoil (*Lotus corniculatus*) shown to reduce the *T. circumcincta* egg output of sheep (Waller *et al.*, 2004b).

1.1.5.4 Nutritional supplementation

Nutritional supplementation with urea or additional protein enhances the immune response to *T. circumcincta* (Stear *et al.*, 2007). Additionally, supplementation with trace elements such as iron, zinc, copper and molybdenum have been shown to influence host resistance to nematode

infection (Koski *et al.*, 2003). Nutritional supplementation is a successful method for nematode control but is usually limited by financial costs.

1.1.5.5 Vaccines

An effective vaccine is the ideal control method for *T. circumcincta*. Irradiated larval vaccines have been successful against some nematode species such as bovine and ovine lungworm (*Dictyocaulus spp.*), however, they do not always generate immunity to gastrointestinal nematodes (Bain, 1999). Irradiated *H. contortus* larvae were shown to be useful against adult sheep but not lambs (Urquhart *et al.*, 1966).

Vaccines against parasite molecules that are recognised by the host during natural or deliberate infection have induced high immune responses, but no single molecule has been shown to produce a high level of resistance (Stear *et al.*, 2007). Vaccines which utilise proteins found in the nematode gastrointestinal tract have shown some success against nematode species that are blood feeders such as *H. contortus*. These proteins do not generally interact with host cells and would not normally induce an immune response, however, the surface of the parasite intestine is exposed to host antibodies through feeding on blood (Knox *et al.*, 2003). Extraction of sufficient amounts of these nematode gastrointestinal proteins is demanding, but recombinant proteins may overcome this issue. A vaccine against *T. circumcincta* developed by Nisbet *et al.* (2013) uses recombinant larval antigens that have been shown to be targets of IgA antibodies from immune sheep. This vaccine is made up of 8 different proteins and has had variable success.

1.1.5.6 Breeding

Selective breeding has also been used to develop resistance against *T. circumcincta* and other nematode species. Different breeds of sheep have been shown to have different resistance levels to nematode species (Bahirathan *et al.*, 1996; Gamble *et al.*, 1992; Nguti *et al.*, 2003; Yazwinski *et al.*, 1979). For example, Texel breed sheep are more resistant to *T. circumcincta* than Suffolk breed sheep (Good *et al.*, 2006). In India, Garole breed sheep are known to have increased resistance to gastrointestinal nematodes, but cross-breeding Deccani and Bannur breeds with Garole was shown to significantly negatively impact live weight and growth rates of lambs (Nimbkar *et al.*, 2003). In Brazil, cross-breeding with Santa Ines breed sheep has been shown to increase production and maintain nematode resistance, especially against *H. contortus* (Amarante *et al.*, 2009). Generally, crossbreeding increases growth above the average of the parents because of heterosis, however, the resistant parent may be quite small. Crossbreeding sheep may be an option for enhancing resistance but needs further investigation.

If crossbreeding is not an option, then selection from within a breed can be done instead. Choosing a set number of characteristics and breeding with sheep that consistently display those

characteristics may help with resistance breeding. Heritability studies have been conducted to determine breedable resistance traits (Bishop *et al.*, 1996; Bisset *et al.*, 2011; Coltman *et al.*, 2001; Schwaiger *et al.*, 1995), and Stear *et al.* (2007) provides an excellent review of what is involved.

All these methods have been used with varying success (W. D. Smith *et al.*, 1986; W. D. Smith *et al.*, 1983, 1985; Stear *et al.*, 1997; Stear *et al.*, 2007). A combination of control measures is likely to provide the most effective and sustainable control and will be dependent upon the farm of interest (Stear *et al.*, 2000). The most appropriate combination will differ between farms and their needs. *T. circumcincta* is a complex species of great importance. Ongoing research in many areas will be required to understand and control this parasite.

1.2 Cysteine Proteases

Cysteine proteases (also known as peptidases or proteinases) are enzymes influencing processes involving cell death, protein degradation, post-translational modifications of proteins, extracellular matrix remodelling, autophagy, and immune signalling (Dana *et al.*, 2020). Parasitic cysteine proteases are involved in parasite stage transition, invasion of host tissues, nutrient uptake, and immune evasion. In helminth parasites, cysteine proteases are most often secreted externally (T. H. Kang *et al.*, 2004).

The cysteine proteases are grouped into clans. Clan CA contains papain, calpain and viral cysteine proteases. The papain (C1) family of clan CA is the largest family, where most cysteine proteases belong (T. H. Kang *et al.*, 2004). The papain-like fold is composed of 2 domains; a left domain and a right domain that form the active protease with the active site cleft between them (Vidak *et al.*, 2019). C1 proteases are characterised by an active site composed of a “catalytic triad”; a cysteine, a histidine and an asparagine residue (Deussing *et al.*, 2000), with the asparagine residue orientating the imidazole ring of histidine to form the catalytic triad (Barrett *et al.*, 2001).

The C1 family consists of two major subfamilies: cathepsin B and cathepsin L. The former is characterised by the presence of an occluding peptide loop, while the latter is characterised by the presence of an ERFNIN motif and comprises cathepsins L, V, K, S, W, F and H (T. H. Kang *et al.*, 2004; Turk *et al.*, 2012). Cathepsins B, C, F, H, L, O and Z are found universally amongst tissues and cell types, whereas cathepsins J, K, L2, S and W are specific to tissues or cell types (Deussing *et al.*, 2000).

Cathepsins are synthesised as zymogens (an inactive precursor) with an N-terminal propeptide which inhibits the enzyme action. Cathepsin propeptide inhibitors are α -helical domains which physically prevent access to the substrate-binding active site cleft. Proteolytic cleavage removes the inhibitor and activates the enzyme (Groves *et al.*, 1996). After activation, their proteolytic activity is kept under control by pH, compartmentalisation, and by inhibitors such as cystatins,

stefins, kininogens, thyropins, and serpins (Barrett *et al.*, 2001; Vidak *et al.*, 2019). Cysteine cathepsins require reducing and mildly acidic conditions for optimal activity, and all (except cathepsin S) are irreversibly inactivated at a neutral pH (Vidak *et al.*, 2019).

Most cathepsins are endopeptidases, however, some cathepsins have exopeptidase activity due to loops and propeptide regions that limit the accessibility of the active site (Vidak *et al.*, 2019). Exopeptidases break terminal amino acid peptide bonds while endopeptidases break non-terminal amino acid peptide bonds.

1.2.1 Cathepsin L

Cathepsin L is involved in invasion and metastasis of tumours, atherosclerosis, renal disease, and viral infection (Y. Y. Li *et al.*, 2017). Cathepsin L, like other cysteine cathepsins, contains an inhibitory propeptide domain that blocks the active site of the mature enzyme preventing premature activation (Dana *et al.*, 2020). Cathepsin L of *Fasciola hepatica*, a trematode species which causes severe liver damage, has been shown to break down cattle IgG antibodies (Carmona *et al.*, 1993; A. M. Smith *et al.*, 1993).

1.2.2 Cathepsin B

Cathepsin B is associated with a variety of human diseases such as tumour metastasis, rheumatoid arthritis, osteoporosis, pancreatitis, inflammatory respiratory disease, and liver fibrosis (Y. Y. Li *et al.*, 2017). Parasitic nematode species have been shown to rely heavily on cathepsin B proteases for larval development and pathogenicity (Duffy *et al.*, 2006). *Radopholus similis*, a parasitic plant nematode of crop species, showed significantly inhibited development and pathogenicity of larvae following silencing of the cathepsin B gene (Y. Li *et al.*, 2015). A cathepsin B homologue of *Parelaphostrongylus tenuis*, a parasitic nematode causing neurological disease in many domestic livestock species, was implicated in larval development and possibly in the emergence of L3 from the intermediate snail host (Duffy *et al.*, 2006).

Unique among the cathepsins, cathepsin B is capable of both exopeptidase and endopeptidase (at neutral pH) activity, due to the presence of a ~20 amino acid insertion termed the “occluding loop” (Illy *et al.*, 1997; Vidak *et al.*, 2019). When cathepsin B acts as an exopeptidase, the occluding loop is held on to the body of the enzyme by two salt bridges blocking the active site cleft, which leaves two histidine residues that bind the substrate’s C-terminal carboxylate enabling the exopeptidase activity (Y. Y. Li *et al.*, 2017).

1.2.3 Cathepsin F

Cathepsin F is a relatively new C1 cysteine protease (T. H. Kang *et al.*, 2004; Sloan *et al.*, 2020; B. Wang *et al.*, 1998). In general, there are three different domains for which the prepropeptide of cathepsin F is composed; a C-terminal peptide similar to cathepsin L, a linker peptide and a N-

terminal segment that is homologous with cystatin (unique in the papain family) (T. H. Kang *et al.*, 2004). It has a long proregion (up to 250 residues) and a highly conserved ERFNAQ motif (in place of the cathepsin L ERFNIN motif). Next to this sequence is an E/DXGTA motif, which has been identified as a proregion feature of cathepsins F and W, but not cathepsin L (Redmond *et al.*, 2006). Cathepsin F proregions also contain a cystatin domain. However, the proregion of *T. circumcincta*, along with *Schistosoma mansoni* and *Clonorchis sinensis*, does not have a cystatin domain (Redmond *et al.*, 2006). *Clonorchis sinensis* instead has four small α -helical domains in the proregion which may be responsible for enzyme stability from autoactivation as the cystatin domain does in other cathepsin F proteases (T. H. Kang *et al.*, 2004).

N-linked glycosylation and signal peptide cleavage occurs during movement to the endoplasmic reticulum. After removal of the signal peptide, the propeptide assists in the folding of the enzyme and targeting to the endosomes/lysosomes using a specific mannose-6-phosphate receptor pathway, while simultaneously acting as an inhibitor to prevent any premature proteolytic activity of the zymogen. After removal of the N-terminal propeptide, the mature catalytically active cathepsin is released (Turk *et al.*, 2012).

Redmond *et al.* (2006) estimated the molecular size of *T. circumcincta* cathepsin F mature domain to be ~24 kDa, which was slightly smaller than the mass of ~40 kDa they calculated from SDS-PAGE. The difference between the two is likely due to glycosylation of the native enzyme. There are two predicted N-glycosylation sites; one on the mature enzyme and one in the pro-region. Glycosylation of the mature protein has been confirmed in lectin-binding affinity experiments (Redmond *et al.*, 2006).

A secreted cathepsin F of *T. circumcincta* is regularly referred to as Tci-CF-1 (GenBank accession: ABA01328) and was isolated from L4 from experimentally infected sheep (Redmond *et al.*, 2006). The function of cathepsin F in *T. circumcincta* has not yet been determined.

1.3 DNA extraction

To look more closely at a particular gene, DNA must be extracted from the organism of interest. Three types of DNA extraction are commonly used: chelating, precipitation, and silica binding. Systematic comparison of various methods determines which type is most appropriate for a particular species or sample type, as all three methods have their place in DNA extraction.

A variety of different methods have been used for DNA extraction of nematode species. Doyle *et al.* (2019) compared 5 commercial kits to extract individual nematode DNA from 8 different species and found a Cancer Genome Project method was ideal. This method utilised Whatman® FTA® cards for sample collection. Seesao *et al.* (2014) compared 4 methods for extraction of pooled Anisakidae nematodes and found a silica binding column was the best method because it

provided good quality and quantity DNA repeatedly and at low cost, however, modifications to the protocols to breakdown the complex nematode cuticle were required.

The most common methods for DNA extraction from *T. circumcincta* have been precipitation and silica binding. Precipitation methods have been used on individual *T. circumcincta* specimens (Gasser *et al.*, 1993; Stevenson *et al.*, 1996), and were used for the current *T. circumcincta* reference genome (Choi *et al.*, 2017). Silica binding column methods have been used on both pooled and individual *T. circumcincta* specimens many times (Ashrafi *et al.*, 2020; Bott *et al.*, 2009; Martinez-Valladares *et al.*, 2020). There is currently no consensus on which DNA extraction method is best for *T. circumcincta*.

1.3.1 Chelating

Chelex™ is a chelating ion-exchange resin made of styrene divinylbenzene copolymers containing paired iminodiacetate ions that bind polar components of cells leading to disruption of cell membranes, and cell lysis. The remaining non-polar DNA is retained in the aqueous solution above the Chelex™. The resin prevents DNA degradation by chelating metal ions that catalyse the breakdown of DNA (Lienhard *et al.*, 2019; Walsh *et al.*, 1991). This method is simple and very fast at ~30 minutes.

1.3.2 Precipitation

Precipitation extraction begins with mechanical disruption of a sample to break open cells. This can be done with a mortar and pestle, cutting the sample into small pieces or with a tissue homogeniser. Cell lysis using heat, detergents or enzymes such as proteinase K can also be used to disrupt the cell membrane, dissolve cellular proteins and free DNA (Elkins, 2012). Precipitation of nucleic acids separates the DNA from cellular debris. Salts neutralise the negative charge on DNA molecules, making them more stable and less water soluble. Ethanol is added to force the precipitation of nucleic acids out of solution. Excess solution is removed, the DNA is washed in ethanol to remove any remaining contaminants, and the DNA is re-dissolved in water for use in downstream applications (Elkins, 2012).

1.3.3 Silica binding

Silica binding methods bind DNA to silica gel or beads in the presence of chaotropic salts and under certain pH conditions (Karp *et al.*, 1998). The salts disrupt the hydrogen bonds between strands and cause the nucleic acids to become hydrophobic, thus binding them to the silica. The remaining debris is washed out using ethanol and a variety of buffers (Elkins, 2012). The result is a pure elution of DNA.

1.3.4 DNA analysis

Spectrophotometric absorbance at different wavelengths helps determine possible contaminants in a DNA sample as well as DNA concentration (Matlock, 2015). DNA absorbs light at 260 nm, however, it does not distinguish between double- and single-stranded DNA, RNA, and nucleotides. Furthermore, impurities such as protein, phenol and other salts may also measure readings at this wavelength. To account for this, the purity of DNA relative to contaminants can be determined by measuring the ratio of different wavelengths, for example, A260/280 and A260/230 (Matlock, 2015). The A260/280 ratio for spectrophotometry is used to determine the presence of protein in a sample, and a pure DNA A260/280 ratio should be 1.8. Lower ratios indicate protein contamination. The A260/230 ratio is used to indicate the presence of organic contaminants which could affect downstream applications. A pure DNA sample should have an A260/230 ratio of 2.0. These ratios are used together to determine the purity of a DNA sample (Matlock, 2015).

Fluorometric analyses are also performed to determine DNA concentration. A fluorescent dye binds specifically to the nucleic acids within a sample and the DNA is quantified by the fluorescence measured by the detector. Even if the sample is contaminated, it can give an accurate reading because the dye is bound only to DNA. Fluorometric analysis is highly regarded for use in sequencing and PCR because it can quantify DNA as low as 10 pg – 200 ng and is considered a very accurate quantification method, however, an additional instrument is required to measure the quality of the sample (Simbolo *et al.*, 2013).

1.4 Genome sequencing

Once DNA has been extracted, DNA sequencing, the process of determining the nucleic acid sequence or order of nucleotides, can begin. The first two methods for DNA sequencing were described in 1977. The first, by Maxam *et al.* (1977), used a chemical breakage, radioisotope labelling and gel electrophoresis method, and the second, by Sanger *et al.* (1977), used dideoxy nucleotide analogues as specific chain-terminating inhibitors of DNA polymerase. Knowledge of DNA sequences has become crucial for basic biological research, and in many applied fields such as biotechnology, medical diagnosis, forensic biology, biological systematics, and virology. Comparing mutated and healthy DNA sequences can diagnose diseases and various cancers (Chmielecki *et al.*, 2014), characterize antibody repertoire (Abate *et al.*, 2013), and can be used to guide patient treatment (Pekin *et al.*, 2011). The current draft genome for *T. circumcincta* was developed using drug-susceptible strains to compare against drug-resistant strains and identify potential drug-resistance genes or mutations (Choi *et al.*, 2017).

Whole genome sequencing is the process of determining the complete DNA sequence of an organism's genome by sequencing all an organism's chromosomal, mitochondrial, and additionally for plants, chloroplast DNA. Genome sequencing improves the knowledge available

to researchers and is a valuable tool for predicting disease susceptibility and drug response (Behjati *et al.*, 2013). Genome sequencing allows researchers to compare the DNA of different organisms and identify their unique evolutionary paths. Quick methods to sequence DNA allows for more organisms to be identified and catalogued (Abate *et al.*, 2013).

1.4.1 Sequencing methods

The three DNA and genome sequencing companies most relevant to this thesis are Illumina, Pacific Biosciences (PacBio) and Oxford Nanopore Technologies.

1.4.1.1 Illumina

In the Illumina sequencing method, sample DNA is fragmented into 75 – 400 bp pieces. Custom adapters are added to the DNA pieces and the library flows across a solid surface known as the flow cell. Prepared sample DNA fragments bind to the solid surface and due to the dense lawn of adaptor complementary sequences on the surface, each will anneal to a nearby primer. This process is followed by bridge amplification where the strands undergo elongation and form a double-stranded bridge on the surface of the flow cell (Su *et al.*, 2011). Denaturation frees the two strands, both now fixed on the flow cell surface at one end and the cycle can repeat. Repeated cycles will form colony-like local clusters with approximately one million copies of each template (Su *et al.*, 2011). The sequencing reaction begins with the addition of a universal sequencing primer, which hybridizes to the adaptor sequences added in the first stage (Liu *et al.*, 2012). Modified dNTPs containing a terminator, which block further polymerization, are used. The terminator also contains a fluorescent label, which can be detected by a camera, can be computationally converted into sequence reads (Quail *et al.*, 2012). One base is incorporated and interrogated at a time since further elongation of the chain is prevented (Bentley, 2006).

When all colonies are scanned at the end of a cycle and the base determined for each colony, the fluorophores are cleaved off and terminating bases are activated, allowing for another round of nucleotide incorporation. The sequencing reaction is conducted simultaneously on a very large number of different template molecules spread out on the flow cell (Moorthie *et al.*, 2011; Pillai *et al.*, 2017).

Illumina reads are considered “short reads” and are a challenge for *de novo* assembly of large, repetitive genomes because of insufficient overlap between the fragments (Adewale, 2020). Substitutions of nucleotides are the most common error type for Illumina sequencing, however, it has a low error rate of approximately 1% (Kchouk *et al.*, 2017).

1.4.1.2 PacBio

PacBio developed the single-molecule real-time (SMRT) sequencing, a parallelized single molecule DNA sequencing method that enables direct observation in real time. Several thousand

nanophotonic visualization chambers called zero-mode waveguides (ZMW) make up the so-called SMRT cell (Levene *et al.*, 2003). Each ZMW is a small chamber, about 70 nm in diameter, constructed in a thin metal film about 100 nm deposited on a glass substrate, where a polymerase is affixed at the bottom with a single molecule of template DNA. The extremely small size of the ZMW prevents visible laser light from passing through. The bottom of the ZMW is only illuminated when a light source (laser) is applied to it. Pacific Biosciences uses four fluorescently labelled nucleotides, which generate distinct emission spectrums. The phospho-linked dNTPs may diffuse in and out of the ZMW. When the affixed polymerase encounters the correct dNTP, the dNTP becomes incorporated, and the fluorophore excited by the laser emits a detectable light signal. The dNTP then cleaves the phosphodiester bond, and the fluorescent tag diffuses out of the observation area of the ZMW where its fluorescence is no longer observable. A detector detects the fluorescent signal of the nucleotide incorporation, and the base call is made according to the corresponding fluorescence of the dye (Eid *et al.*, 2009; Quail *et al.*, 2012; Vilgis *et al.*, 2018).

PacBio sequencing produces “long reads” which overcomes the overlap issue seen with short reads. Greater sequence overlap allows for better sequence assembly and accuracy of repeat regions (Kchouk *et al.*, 2017). PacBio readily produces reads of ~10 kbp, but has been shown to produce reads of >200 kbp in length (Kraft *et al.*, 2019). Unfortunately, PacBio had a high error rate of about 13%, mostly insertions and deletions, and randomly distributed along a read (Kchouk *et al.*, 2017). In more recent years this error rate has been reduced to <1% (Amarasinghe *et al.*, 2020).

1.4.1.3 Oxford Nanopore

A nanopore is a nano-scale hole with a size of about 1.4 nm diameter (Liu *et al.*, 2012). Holes can be created by proteins puncturing membranes (biological nanopores) or in solid materials (solid-state nanopores) (Vilgis *et al.*, 2018). A protein nanopore is set in an electrically resistant polymer membrane. The membrane is immersed in an electrolyte solution, and an ionic current is passed through the nanopore by setting a voltage across the membrane. Single molecules entering the nanopore cause characteristic disruptions in the current, and by measuring this disruption, DNA or RNA molecules can be identified and characterised (Kasianowicz *et al.*, 1996; Vilgis *et al.*, 2018).

Oxford Nanopore Technology (ONT) sequencers are currently able to produce the longest reads on the market, at upwards of 2 Mbp, and are often referred to as “ultra-long reads” (Kraft *et al.*, 2019). Similar to PacBio, ONT were known to have high error rates of ~12%, distributed amongst mismatches, insertions and deletions (Kchouk *et al.*, 2017), but has dropped to ~5% in recent years (Amarasinghe *et al.*, 2020).

1.4.2 *Teladorsagia circumcincta* genome

Caenorhabditis elegans is the model organism for nematode species and has been extensively researched. *C. elegans* is in Order Rhabditida alongside *T. circumcincta*, *Necator americanus*, and *H. contortus*, among others. The genome of *C. elegans* is 100 Mb (Wood, 2002a), with 19,735 protein-coding genes (Hillier *et al.*, 2005). *N. americanus* has an estimated genome size of 244 Mb, and 19,151 protein-coding genes (Tang *et al.*, 2014). *H. contortus*, one of *T. circumcincta*'s closest relatives, has an estimated genome size of 320 – 370 Mb, and 21,732 – 23,610 protein-coding genes (Laing *et al.*, 2013; Schwarz *et al.*, 2013). The size of a nematode genome can vary considerably, however, none of these genomes are complete and final sizes are still to be determined.

A draft genome for *T. circumcincta* was developed by Choi *et al.* (2017) using partially in-bred drug-susceptible strains. This draft genome was developed to compare genome-wide single nucleotide and copy number variants of drug-resistant *T. circumcincta* strains. The estimated *T. circumcincta* genome size from this assembly was ~700 Mb, comprised 81,730 contigs, and has an estimated 25,532 protein-coding genes (Choi *et al.*, 2017). This draft was sequenced using a Genome Sequencer Titanium FLX (Roche Diagnostics, Basel, Switzerland) which generates short reads 330 – 500 bp in length, as well as Illumina paired-end reads ranging 75 – 150 bp in length (Kchouk *et al.*, 2017). Short read sequences generally have lower sequencing error rates (Dominguez Del Angel *et al.*, 2018; Kchouk *et al.*, 2017; Quail *et al.*, 2012), however, they often fail to generate sufficient overlap in the DNA fragments causing major problems for *de novo* assembly. Discontinuous contigs or large repetitive regions may occur because PCR preferentially amplifies repetitive DNA (Adewale, 2020).

The Wellcome Trust Sanger Institute have also done some *T. circumcincta* genome sequencing (BioProject PRJEB7676), using predominantly PacBio SMRT sequencing and some Illumina HiSeq sequencing on pooled L3 *T. circumcincta*, totalling 240 Gbp of data, all of which is available through the Sequence Read Archive. There is no indication that this data has been used in an assembly as yet.

Flow cytometry has been considered a fast, sensitive technique for determining genome sizes in many organisms (Nath *et al.*, 2014; J. Wang *et al.*, 2015). Leroy *et al.* (2003) used flow cytometry to estimate genome sizes of a range of nematode species, using the known *C. elegans* genome as calibration. Estimated genome sizes for *H. contortus* and *T. circumcincta* were calculated at 52.6 and 58.6 Mb, respectively. This is significantly different from the 700 Mb genome size estimated in the Choi *et al.* (2017) genome assembly and requires further investigation.

1.5 Research Plan

To adequately study the functions of proteins used by *T. circumcincta* to evade the immune system or feed on the mucosal surface, we need a thorough understanding of its genome. This thesis aims to identify and uncover what is known about *T. circumcincta* cathepsin F and how it fits into the grand scheme of parasite function. Specifically, this thesis aims to

1. interrogate the available draft *T. circumcincta* genome and find the gene responsible for cathepsin F, model the protein structure produced by that gene, and compare the cathepsin F protein and gene to cathepsins of closely related nematode species,
2. determine an appropriate DNA extraction method for high quality genome sequencing using long-read technology,
3. generate a draft genome using long-read sequencing technology to explore the cathepsin F gene, and
4. to create recombinant *T. circumcincta* cathepsin F protein *in vitro* for diagnostic and characterisation purposes.

Chapter 2

Cathepsin F of *Teladorsagia circumcincta* is a recently evolved cysteine protease

2.1 Chapter Preface

Cathepsin F of *Teladorsagia circumcincta* has been previously identified as a target of IgA antibodies from immune sheep and is used in a multivalent recombinant vaccine against Teladorsagiosis. However, little is known about its structure or function. This chapter represents the first published manuscript of this thesis and presents the bioinformatic analysis of the cathepsin F structure and gene. The protein sequence and structure, and gene of cathepsin F were evaluated to find species homologues, determine its phylogenetic history and potential functional role.

This study found that *T. circumcincta* cathepsin F is a recently evolved cysteine protease that does not fall clearly into either of the cathepsin L or F subfamilies. It exhibits characteristics of both cathepsins F and L, and the sequence similarity to its closest homologues is low, including proteins of closely related nematodes of the same subfamily.

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 (A. M. Smith *et al.*, 1993). Cathepsin F could be digesting host IgA for protection against the host immune response or as a source of protein to aid in development of the nematode. The information collected in this study is intended to inform the design of further functional studies.

This chapter is presented in published format.

2.2 Publication details

Title: Cathepsin F of *Teladorsagia circumcincta* is a recently evolved cysteine protease

Journal details: Evolutionary Bioinformatics, 2020, September 2, doi:

<https://doi.org/10.1177/1176934320962521>

Stage of publication: Published

Authors: Sarah Sloan, Caitlin Jenvey, Callum Cairns and Michael Stear

2.3 Statement of contribution of joint authorship

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.

2.4 Statement from the co-author confirming the authorship contribution of the PhD candidate

“As co-author of the manuscript ‘Sloan, S., Jenvey, C., Cairns, C., & Stear, M. (2020). Cathepsin F of *Teladorsagia circumcincta* is a recently evolved cysteine protease. *Evol Bioinform Online*, 16, 1176934320962521. doi:10.1177/1176934320962521’ I can confirm that Sarah Sloan made the following contributions:

- Literature review
- Development of the experimental design
- Annotation of protein sequences
- Sequence data extraction and analysis
- Gene construction
- Polymorphism identification
- Multiple sequence alignment of homologous genes in related species
- Pairwise comparisons of homologous genes in related species
- Phylogenetic tree construction and analysis
- Phylogenetic network construction and analysis
- Homology modelling of secondary and tertiary structures
- Generated all figures
- Writing the manuscript, critical appraisal of content and response to reviewers”

Date: 19/09/2020

Cathepsin F of *Teladorsagia circumcincta* is a recently evolved cysteine protease

Sarah Sloan¹, 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 bovines 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 fourth-stage 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/ham/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/s-andrews/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 mis-assembly 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	MSLLFLLIIP	HLFAATVKQQ	YSGGVKPLTE	LRTDLIDKKT	KGSIEFARLG	QHISPKDFGA
Nisbet et al. (2013) polymorph		*.T.....
Variant 1	
Variant 2	
Variant 3	
	Tci-CF-1	WNHFTSFIER	HDKVYRNESE	ALKRFGIFKR	NLEIIRSAQE	NDKGTAIYGI	NQFADLSPEE
Nisbet et al. (2013) polymorph		D.....
Variant 1	
Variant 2	
Variant 3	
	Tci-CF-1	FKKTHLPHTW	KQPDHPNRIV	DLAEGVDPK	EPLPESFDWR	EHGAVTKVKT	EGHCAACWAF
Nisbet et al. (2013) polymorph	
Variant 1	
Variant 2	
Variant 3	
	Tci-CF-1	SVTGNIEGQW	FLAKKKLVSL	SAQQLLDCDV	VDEGCNGGFP	LDAYKEIVRM	GGLEPEDKYP
Nisbet et al. (2013) polymorph	
Variant 1	
Variant 2	
Variant 3	
	Tci-CF-1	YEAKAEQCRL	VPSDIAVYIN	GSVELPHDEE	KMRAWLVKKG	PISIGITVDD	IQFYKGGVSR
Nisbet et al. (2013) polymorph	
Variant 1	
Variant 2	
Variant 3	
	Tci-CF-1	PTTCRLSSMI	HGALLVGYGV	EKNIPYWIHK	NSWGPNGWGED	GYRMRVRGEN	ACRINRFPTS
Nisbet et al. (2013) polymorph	
Variant 1	
Variant 2	
Variant 3	
	Tci-CF-1	AVVL					
Nisbet et al. (2013) polymorph						
Variant 1						
Variant 2						
Variant 3						

Figure 1. Multiple sequence alignment of *Teladorsagia circumcincta* secreted cathepsin F (Tci-CF-1, GenBank accession no. ABA01328), the polymorphism discussed in Nisbet et al.¹³ 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 al.¹³ 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 pseudo-gene 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,

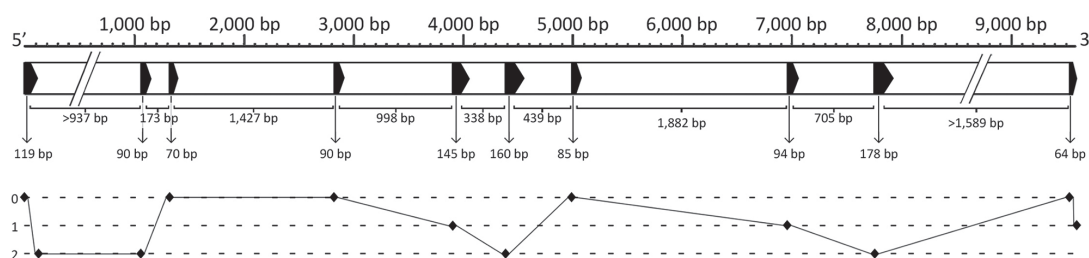


Figure 2. *Teladorsagia circumcincta* cathepsin F variant 3 gene exon/intron structure, constructed using PacBio and Illumina reads in SPAdes and CLC. Exons indicated by solid black arrows, introns indicated by white segments, gaps in contigs indicated by a break. The positions of intron-exon junctions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

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_010213 (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

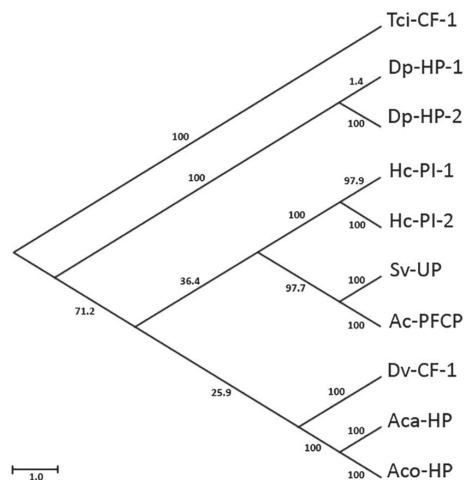


Figure 3. Phylogenetic tree of Tci-CF-1 and its 9 closest homologs. Tci-CF-1: *Teladorsagia circumcincta* secreted cathepsin F (GenBank accession no. ABA01328); Dp-HP-1: *Diploscapter pachys* hypothetical protein WR25_25536 (GenBank accession no. PAV60527); Hc-PI-1: *Haemonchus contortus* proteinase 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 E₈₀R₈₄F₈₈N₉₁A₉₅Q₉₉, in which alanine (A₉₅) has been substituted for isoleucine (I₉₅), E/D₁₀₂xG₁₀₄T₁₀₅A₁₀₆, and G₁₀₉xN₁₁₁xF₁₁₃xD₁₁₅ (Figure 6). The predicted N-terminal of the mature processed protein is located at residue E₁₅₁ (Figure 5). The mature protein contains the conserved structural motif G₂₁₄C₂₁₅N₂₁₆G₂₁₇G₂₁₈ as well as the catalytic triad active site residues C₁₇₇, H₃₁₁, and N₃₃₁. Two predicted N-glycosylation sites were found at residues N₇₇E₇₈S₇₉ and N₂₆₀G₂₆₁S₂₆₂ (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 (E₃₀ > D₃₀), 56 (K₅₆ > N₅₆), 76 (R₇₆ > K₇₆), 235 (P₂₃₅ > S₂₃₅), and 306 (L₃₀₆ > P₃₀₆) (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 amino acid 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

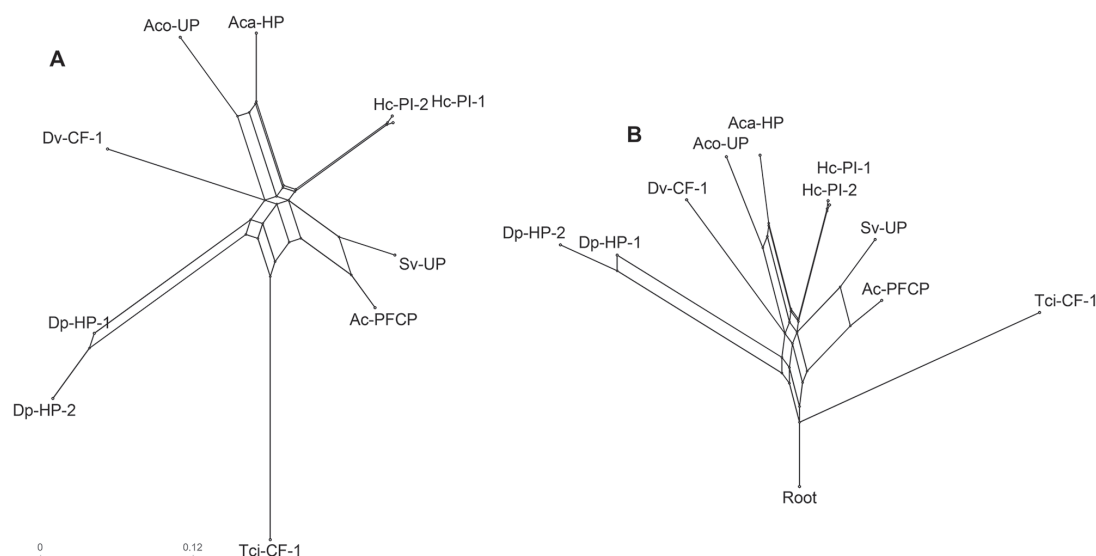


Figure 4. Phylogenetic networks of Tci-CF-1 and its 9 closest homologues; Tci-CF-1: *Teladorsagia circumcincta* secreted cathepsin F (GenBank accession no. ABA01328); Dp-HP-1: *Diploscapter pachys* hypothetical protein WR25_25536 (GenBank accession no. PAV60527); Hc-PI-1: *Haemonchus contortus* proteinase 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. 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 al^{47–49} 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 (P₂₃₅) to serine (S₂₃₅), and leucine (L₃₀₆) to P₃₀₆), as these are likely to affect mature protein function and/or recognition. The substitution of P₂₃₅ to S₂₃₅ 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 L₃₀₆ 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

```

ATGTCCTCTTTGTTCTCCTGCTTCTCATCCACATCTATTGCGGCTACTGTAAGCAGCAATACTCAGGAGGTGTCAAACCGTTGACA
M S L L F L L L I P H L F A A T V K Q Q Y S G G V K P L T

GAATTGCGTACGGATTGATCGACAAGAAGACCAAAGGCTCGATCGAGTTCGCCAGGCTTGGTCAACACATCAGTCCAAAAGACTTC
E L R T D L I D K K T K G S I E F A R L G Q H I S P K D F

GGTGCATGGAATCATTTACACAGCTTCATTGAAAGGCATGACAAGGTCTACAGAAACGAGAGCGAAGCTCTGAAACGATTGGGATC
G A W N H F T S F I E R H D K V Y R N E S E A L K R F G I

TTCAAGAGAAATCTCGAGATAATTCGCTCTGCGCAGGAAACGATAAGGGAACAGCTATTTACGGAATCAATCAGTTTGCTGATCTT
F K R N L E I I R S A Q E N D K G T A I Y G I N Q F A D L

TCACCGGAGGAATTCAAAAGACTCACCTGCCGCACACATGGAACAGCCTGATCATCAAACCGAATCGTGGACTTAGCCGCAGAA
S P E E F K K T H L P H T W K Q P D H P N R I V D L A A E

↓
GGGGTGGATCCGAAGGAGCCACTGCCGAATCGTTTCGATTGGAGAGAACATGGTGCAGTGACAAAAGTGAAGGTCAGTGT
G V D P K E P L P E S F D W R E H G A V T K V K T E G H C

GCAGCCTGCTGGGCATTTTCTGTACAGGAAATATGAAGGCCAGTGGTTCCTTGCCAAAAAGAACTTGATCGCTCTCGGCACAA
A A C W A F S V T G N I E G Q W F L A K K K L V S L S A Q

CAGCTCCTCGATTGTGATGTTGTTGATGAGGGATGTAACGGTGGATTTCCTCTTGACGCTTACAAAGAAATCGTTTCAATGGCGGC
Q L L D C D V V D E G C N G G F P L D A Y K E I V R M G G

TTGAACAGAGAAGACAAGTATCCCTACGAAGCCAAGGCAGAGCAGTGTGCCTTGTCCTCATCGGATATCGTGTATATCAACGGC
L E P E D K Y P Y E A K A E Q C R L V P S D I A V Y I N G

TCAGTCGAGCTACCACATGATGAAGAAAAATGAGGGCATGGCTAGTGAAGAAGGGGCCGATATCGATAGGTATCACCGTAGATGAC
S V E L P H D E E K M R A W L V K K G P I S I G I T V D D

ATACAGTTCTATAAAGCGCGCTTTCTCGTCCGACTACCTGTAGACTATCTTCTATGATTTCATGGCGCTCTCTCTGGTCGGATACGGT
I Q F Y K G G V S R P T T C R L S S M I H G A L L V G Y G

GTCGAGAAGAATATACCGTACTGGATTATAAAGAAATTCGTGGGGCCCCAATTGGGGAGAGGATGGATATTACAGGATGGTGCCTGGG
V E K N I P Y W I I K N S W G P N W G E D G Y Y R M V R G

GAGAACGCTTGTCGCATAAACAGATTCCCCAGTCAGCTGTTGTCTCTATAA
E N A C R I N R F P T S A V V L *

```

Figure 5. Annotated *Teladorsagia circumcincta* secreted cathepsin F sequence (GenBank accession no. DQ133568) with annotations. Signal sequence is underlined; predicted N-glycosylation sites marked with squiggle underline; N-terminal amino acid of mature protein indicated by black arrow; catalytic triad active site residues highlighted black; ERFNAQ, E/DxGTA, GxNxExD 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

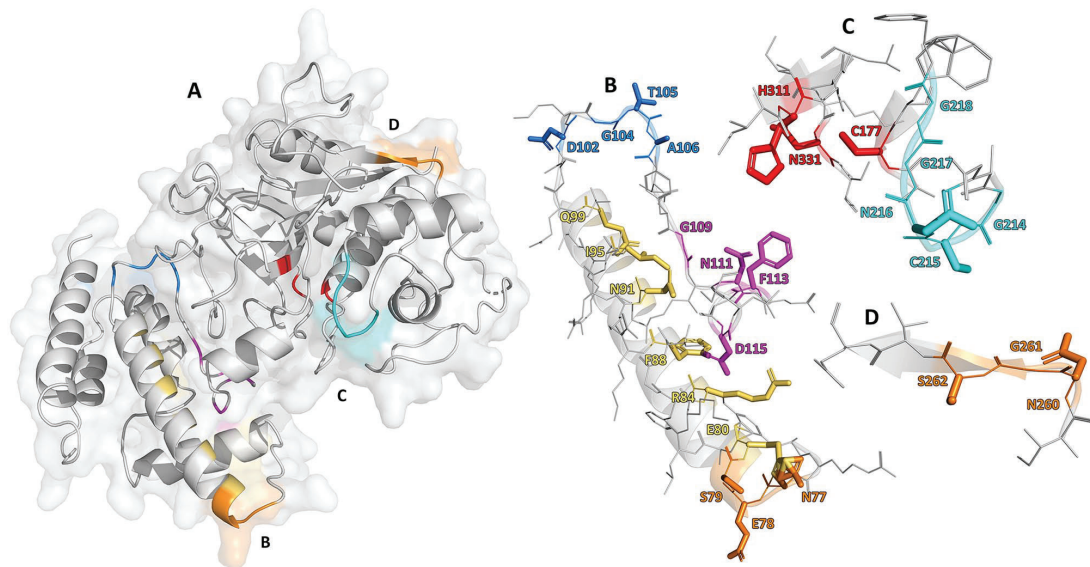


Figure 6. Homology model of *Teladorsagia circumcincta* secreted cathepsin F (GenBank accession no. ABA01328). (A) Ribbon structure showing alpha helices, beta sheets, catalytic triad residues (C, H, N), predicted N-glycosylation sites, ERFNAQ, E/DxGTA, GCNNG, and GxNxFxD motifs. (B) Magnified ERFNAQ,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

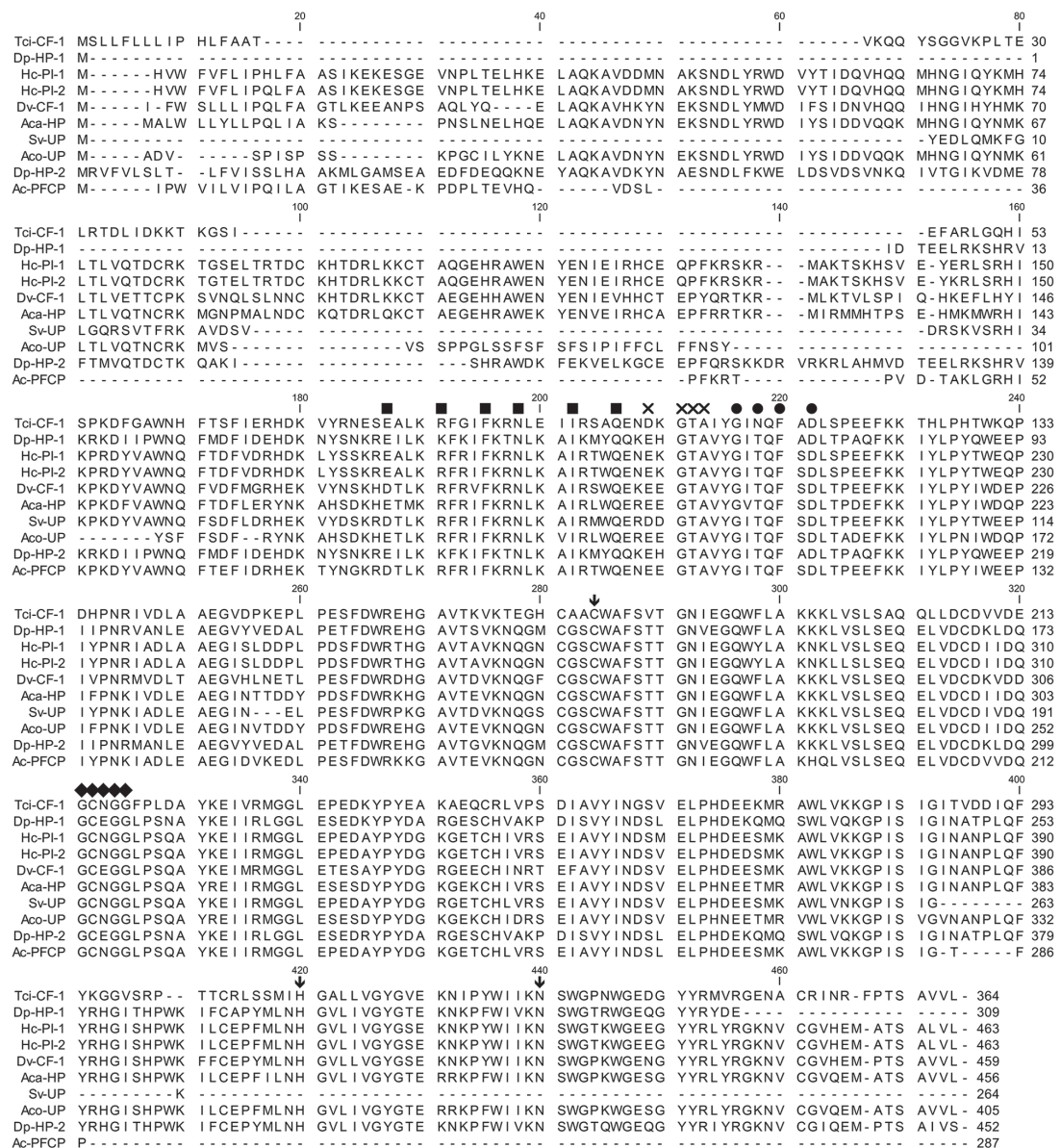


Figure 7. Multiple sequence alignment of Tci-CF-1 with the 9 closest homologous sequences. Tci-CF-1: *Teladorsagia circumcincta* secreted cathepsin F (GenBank accession no. ABA01328); Dp-HP-1: *Diploscapter pachys* hypothetical protein WR25_25536 (GenBank accession no. PAV60527); Hc-PI-1: *Haemonchus 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); Aca-PFCP: *Ancylostoma ceylanicum* papain family cysteine protease (GenBank accession no. EPB70524); Aca-HP: *Angiostrongylus cantonensis* hypothetical protein Angca_010213 (GenBank accession no. KAE9418773); Ss-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 out-group, 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

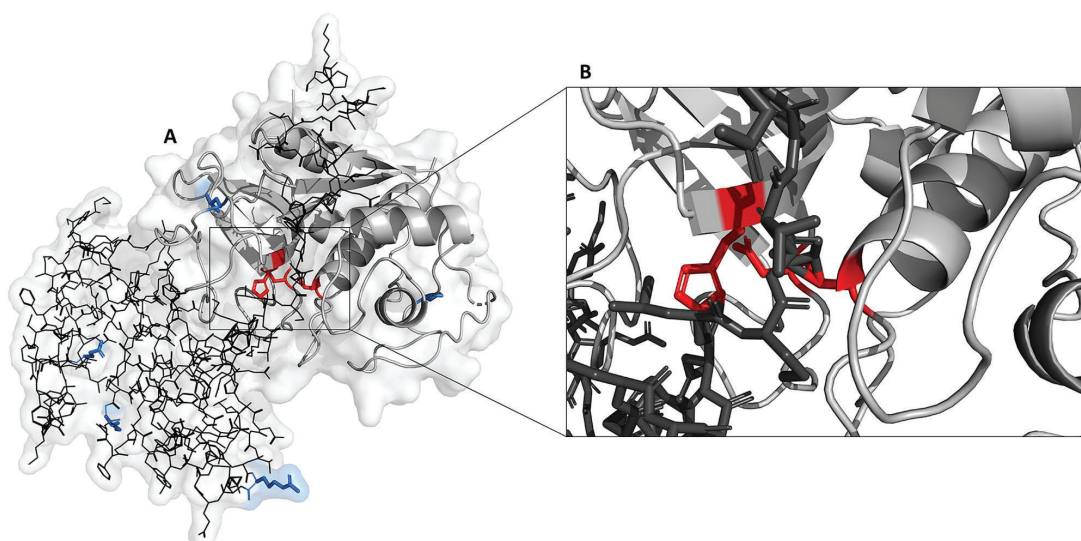


Figure 8. Homology model of *Teladorsagia circumcincta* secreted cathepsin F (GenBank accession no. ABA01328). (A) Pro-region (line residues), mature domain (ribbon), locations of polymorphisms in variants 1, 2, and 3 (bold side-chain residues), and the catalytic triad (bold side-chain residues) which 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.

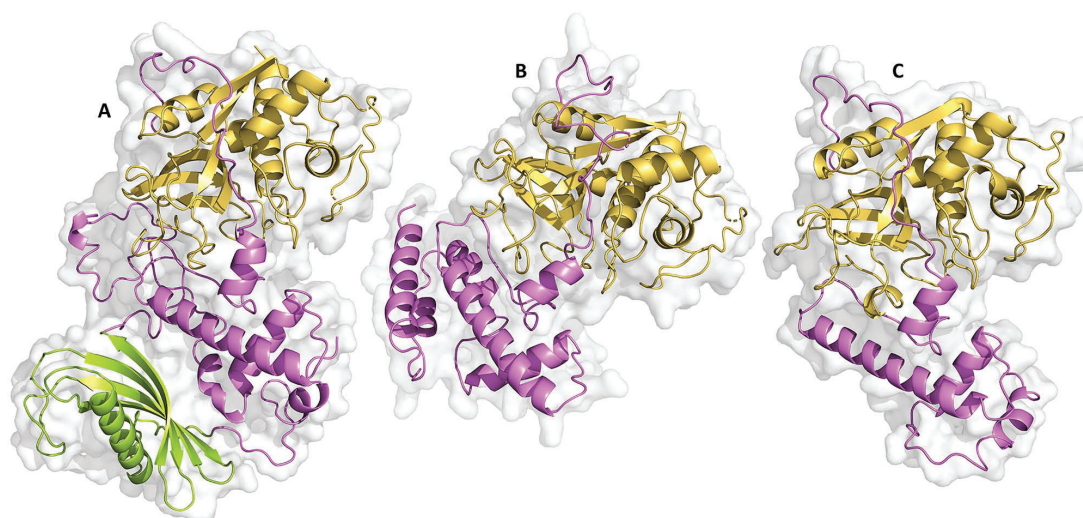


Figure 9. Structural comparison of human and *Teladorsagia circumcincta* cathepsin proteins. X-ray crystallography of mature domains of human cathepsins L⁴⁰ and F⁴¹ and homology modeling of *T. circumcincta* cathepsin F (GenBank accession no. ABA01328) and the pro-regions of human cathepsins L (UniProtKB 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  <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. *Parasitology*. 2015;142(Suppl 1):S26-S39.
- 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.
- Stear MJ, Bairden K, Bishop SC, et al. The genetic basis of resistance to *Ostertagia 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.
- 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 eosinophil 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: propeptide cleavage sites and autoactivation of the zymogen secreted from gastroduodenal 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.

42. Barrett AJ, Rawlings ND. Evolutionary lines of cysteine peptidases. *Biol Chem.* 2001;382:727-733.
43. 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.
44. 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.
45. Stear MJ, Singleton D, Matthews L. An evolutionary perspective on gastrointestinal nematodes of sheep. *J Helminthol.* 2011;85:113-120.
46. Stear MJ, Strain S, Bishop SC. Mechanisms underlying resistance to nematode infection. *Int J Parasitol.* 1999;29:51-56; discussion 73-55.
47. 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.
48. 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.
49. 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.
50. Kolkman JA, Stemmer WPC. Directed evolution of proteins by exon shuffling. *Nat Biotechnol.* 2001;19:423-428.
51. Betts MJ, Russell RB. Amino Acid Properties and Consequences of Substitutions. In: Barnes MR, Gray IC, eds. *Bioinformatics for Geneticists*. Chichester, UK: John Wiley & Sons, Ltd.; 2003:289-316.
52. 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.
53. Parkinson J, Mitreva M, Whitton C, et al. A transcriptomic analysis of the phylum Nematoda. *Nat Genet.* 2004;36:1259-1267.

Chapter 3

Comparative evaluation of different molecular methods for DNA extraction from individual *Teladorsagia circumcincta* nematodes

3.1 Chapter Preface

Many different methods are currently in use for the extraction of *Teladorsagia circumcincta* DNA. The purpose of this study was to develop a reliable DNA extraction protocol for use on individual *T. circumcincta* nematode specimens to produce high quality DNA for genome sequencing and phylogenetic analysis. Pooled samples have been critical in providing the groundwork for *T. circumcincta* genome construction, but there is currently no standard method for extracting high-quality DNA from individual nematodes. With so many commercially available extraction kits on the market there is a need to systematically compare the different methods for optimal extraction of *T. circumcincta* DNA.

Here, comparison of multiple DNA extraction protocols were conducted to determine which methods are most suitable for individual *T. circumcincta* nematode DNA extraction. 11 different DNA extraction protocols on individual *T. circumcincta* nematode specimens were compared based on the yield, quality and reliability of DNA, and protocol time, to obtain DNA suitable for use in PCR and genome sequencing applications.

This study provides the foundation for genome sequencing and assembly in Chapter 4.

This chapter is presented in published format.

3.2 Publication details

Title: Comparative evaluation of different molecular methods for DNA extraction from individual *Teladorsagia circumcincta* nematodes

Journal details: BMC Biotechnology, 2021, May 17, doi:
<https://doi.org/10.1186/s12896-021-00695-6>

Stage of publication: Published

Authors: Sarah Sloan, Caitlin Jenvey, David Piedrafita, Sarah Preston and Michael Stear

3.3 Statement of contribution of joint authorship

SS, CJ, and MS conceived the study. SS designed and performed the research and data analysis, and wrote the manuscript. DP and SP undertook the sampling procedures. CJ and MS verified the analytical methods and supervised the findings of this work. All authors read and approved the final manuscript.

3.4 Statement from the co-author confirming the authorship contribution of the PhD candidate

“As co-author of the manuscript ‘Sloan, S., Jenvey, C. J., Piedrafita, D., Preston, S., & Stear, M. J. (2021). Comparative evaluation of different molecular methods for DNA extraction from individual *Teladorsagia circumcincta* nematodes. *BMC Biotechnol*, 21(1), 35. doi:10.1186/s12896-021-00695-6’ I can confirm that Sarah Sloan made the following contributions:

- Literature review
- Development of the experimental design
- Preparation of *T. circumcincta* specimens
- DNA extraction method selection
- Preparation of DNA extraction reagents
- DNA extractions
- PCR
- Agarose gel electrophoresis
- Sequence data extraction and analysis
- Statistical analysis
- Multiple sequence alignment of PCR products
- Generated all figures
- Writing the manuscript, critical appraisal of content and response to reviewers”

Date: 19/05/2021

RESEARCH ARTICLE

Open Access

Comparative evaluation of different molecular methods for DNA extraction from individual *Teladorsagia circumcincta* nematodes

S. Sloan^{1*}, C. J. Jenvey¹, D. Piedrafita², S. Preston² and M. J. Stear¹

Abstract

Background: The purpose of this study was to develop a reliable DNA extraction protocol to use on individual *Teladorsagia circumcincta* nematode specimens to produce high quality DNA for genome sequencing and phylogenetic analysis. Pooled samples have been critical in providing the groundwork for *T. circumcincta* genome construction, but there is currently no standard method for extracting high-quality DNA from individual nematodes. 11 extraction kits were compared based on DNA quality, yield, and processing time.

Results: 11 extraction protocols were compared, and the concentration and purity of the extracted DNA was quantified. Median DNA concentration among all methods measured on NanoDrop 2000™ ranged between 0.45–11.5 ng/μL, and on Qubit™ ranged between undetectable – 0.962 ng/μL. Median A260/280 ranged between 0.505–3.925, and median A260/230 ranged – 0.005 – 1.545. Larval exsheathment to remove the nematode cuticle negatively impacted DNA concentration and purity.

Conclusions: A *Schistosoma* sp. DNA extraction method was determined as most suitable for individual *T. circumcincta* nematode specimens due to its resulting DNA concentration, purity, and relatively fast processing time.

Keywords: DNA isolation, *Teladorsagia circumcincta*, DNA extraction, Polymerase chain reaction, Nematode, Genome sequencing, Methodology

Background

Parasitic infections of livestock are of major socio-economic importance worldwide. Internal parasites of sheep alone have been shown to cost \$436 million annually in Australia [1]. Therefore, major economic gains are to be made by improving the control of parasitic diseases. Parasitic infection of livestock is largely controlled by anthelmintic treatment, however,

drug resistance is rapidly developing [2–4]. Additional methods of control include nutritional supplementation, vaccination, selective breeding, and pasture management, which are used with varying success [5–9].

Teladorsagia circumcincta is the most important parasitic nematode of sheep in cool temperate regions worldwide [3]. Clinical disease caused by *T. circumcincta* infection results in reduced production, decreased animal welfare, parasitic gastroenteritis, poor growth performance, and weight loss [10]. To develop new control strategies for *T. circumcincta* infection, it is pivotal that research uncovers as

* Correspondence: s.sloan@latrobe.edu.au

¹AgriBio Centre for AgriBioscience, Department of Animal, Plant and Soil Sciences, School of Life Sciences, La Trobe University, 5 Ring Road, Bundoora, Victoria 3086, Australia

Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

much information about the biology of this nematode as possible. An important starting point involves the genomic investigation of *T. circumcincta*.

Nematode species are genetically diverse and identification of variation in genes amongst populations requires the analysis of individual nematodes. Effective extraction of high-quality DNA from individual specimens is essential to assist in developing improved diagnostic methods. *T. circumcincta* are slender, reddish-brown worms with a short buccal cavity [11]. Their size varies considerably amongst sheep, as worm length is affected by host immune pressure [12], but typically female size ranges from 8 to 10 mm, and males 6–8 mm [11]. Due to their small size, there is very limited tissue from which to derive genetic material. Pooled samples have been critical in providing the groundwork for *T. circumcincta* genome construction [13–15], but there is currently no standard method for extracting high-quality DNA from individual nematodes.

With so many commercially available extraction kits on the market there is a need to systematically compare the different methods for optimal extraction of *T. circumcincta* DNA. The method must ensure that adequate *T. circumcincta* DNA is being extracted, as internal parasites are likely to harbour host and host microorganism DNA, in addition to their own, causing complications downstream. The ideal extraction method should optimise DNA quantity, avoid contamination, minimise degradation and inhibitors, require low-cost consumables and equipment, and have rapid processing time. The quantity and purity of DNA are also important for downstream applications, such as PCR or genome sequencing.

Studies focusing on several DNA extraction methods are uncommon. SR Doyle et al. [16] compared 5 methods to extract individual nematode DNA from 8 different species and found a Cancer Genome Project method was ideal. This method utilised Whatman® FTA® cards for sample collection which did not limit the DNA extraction and whole genome sequencing of parasite samples. Y Seesao et al. [17] compared 4 methods for extraction of pooled Anisakidae nematodes and found a silica binding column was the best method because it provided good quality and quantity DNA repeatedly and at low cost, however, modifications to the protocols to breakdown the complex nematode cuticle was required. LM Schiebelhut et al. [18] compared 8 extraction techniques for species comprising 8 separate phyla and found silica binding column methods produced quality DNA quickly, but commercial kits are costly. RL Smith et al. [19] compared 13 extraction methods for ancient powdery mildew specimens and found a silica binding column method was most suitable and that

DNA concentration was more important than quality for whole genome next generation sequencing purposes on limited and valuable specimens.

Three types of DNA extraction have been tested in this study: chelating, precipitation, and silica binding. Chelex™ is a chelating ion-exchange resin that binds polar components of cells leading to disruption of cell membranes, cell lysis and denaturation of DNA. The remaining non-polar DNA is retained in the aqueous solution above the Chelex™ [20, 21]. Precipitation extraction involves salt and ethanol added to an aqueous solution which precipitates nucleic acids. Silica binding methods bind DNA to silica surfaces in the presence of certain salts and under certain pH conditions [22]. All three methods have their place in DNA extraction, with some working better than others in different circumstances, and it is a comparison of various methods that determines which type is most appropriate for a particular species or sample type.

Here we have compared multiple DNA extraction protocols to determine whether silica binding column, precipitation or chelating methods are most suitable for individual *T. circumcincta* nematode DNA extraction. The aim of this study was to compare 11 different DNA extraction protocols on individual *T. circumcincta* nematode specimens based on the yield, quality and reliability of DNA, and protocol time, to obtain DNA suitable for use in PCR and genome sequencing applications.

Results

11 common DNA extraction protocols were compared and selected to encompass a range of extraction methods and modifications, as well as prior availability in the laboratory. These methods included AccuPrep® Genomic DNA Extraction - Mammalian Tissue (AccM), AccuPrep® Genomic DNA Extraction Kit - Whole Blood, Buffy Coat and Cultured Cells (AccW), Chelex®100 (CheX), cetyl trimethyl ammonium bromide (CTAB), E.Z.N.A.® Forensic DNA (EznF), Isolate II Genomic DNA Kit (IsoG), *Schistosoma sp.* DNA Extraction Method (Schi), Schi with larval exsheathment (Schi-LE), sodium dodecyl sulphate (SDS), Wizard® Genomic DNA Purification Kit - Mouse Tail (WizM), and Wizard® Genomic DNA Purification - Plant Tissue (WizP) (Table 1). These protocols were compared based on DNA concentration, quality, purity, and protocol time. The DNA samples were expected to comprise *T. circumcincta* DNA, host DNA from sheep, as well as DNA from microorganisms present in the sheep gut prior to nematode collection. PCR and ITS-2 phylogeny were performed to confirm the protocols would extract *T. circumcincta* DNA.

Table 1 DNA extraction protocols tested on six individual adult, female *T. circumcincta* specimens per method in this study

Method or kit name	Protocol code	Reference or supplier (catalogue #)	Extraction method	Time required (hours)
AccuPrep® Genomic DNA Extraction - Mammalian Tissue	AccM	Bioneer (K-3032)	Silica binding	1.5
AccuPrep® Genomic DNA Extraction Kit - Whole Blood, Buffy Coat and Cultured Cells	AccW	Bioneer (K-3032)	Silica binding	0.5
Chelex®100	CheX	PS Walsh et al. [20]	Chelating	0.5
Cetyl Trimethyl Ammonium Bromide	CTAB	T Sarkinen et al. [23]	Precipitation	6
E.Z.N.A.® Forensic DNA	EznF	Omega Bio-tek (D3591-00)	Silica binding	1.5–2
Isolate II Genomic DNA Kit	IsoG	Bioline (BIO-52066)	Silica binding	1.5–3
<i>Schistosoma</i> sp. DNA Extraction Method	Schi	PJ Brindley et al. [24]	Precipitation	1.5
Sodium Dodecyl Sulphate	SDS	K Edwards et al. [25]	Precipitation	1.25–1.75
Wizard® Genomic DNA Purification Kit - Mouse Tail	WizM	Promega (A1120)	Precipitation	4–4.5 ^a
Wizard® Genomic DNA Purification - Plant Tissue	WizP	Promega (A1120)	Precipitation	1 ^a
Larval exsheathment	-LE	HJ Dawkins et al. [26]	–	1

^a: Additional overnight incubation required**DNA concentration**

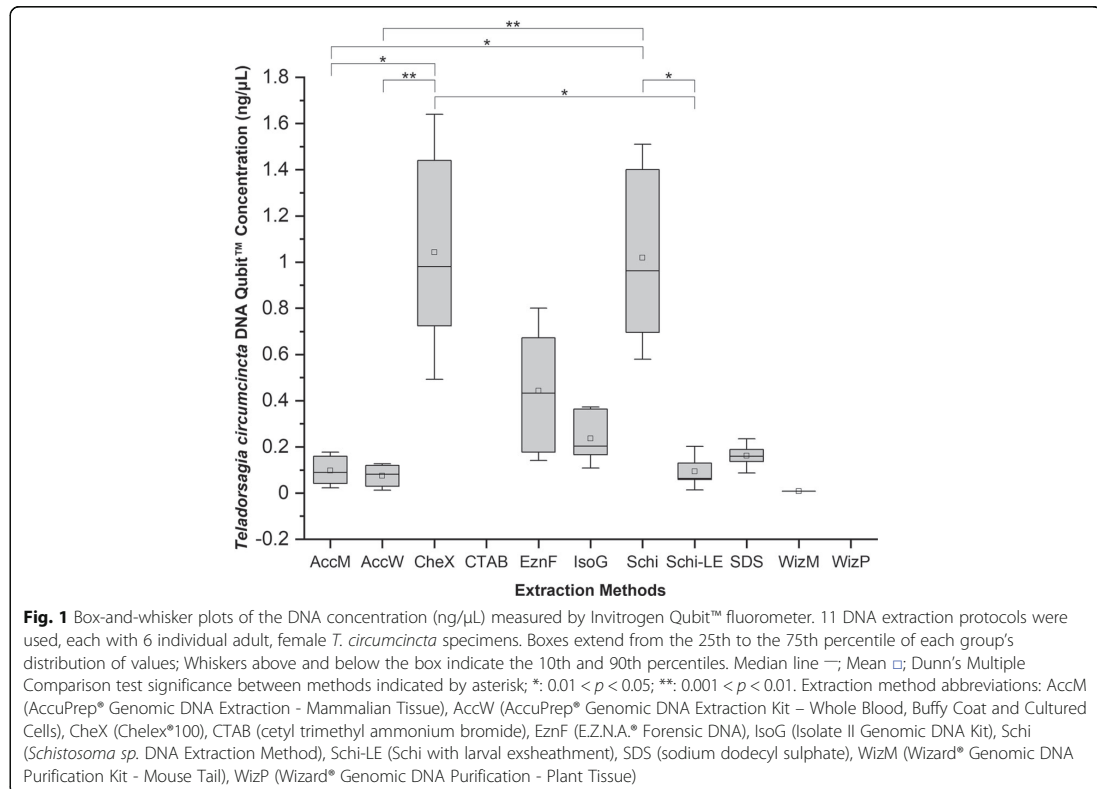
The 11 different DNA extraction protocols generated variable concentrations of DNA from *T. circumcincta* nematode specimens (Table 2). The Qubit™ fluorometer and NanoDrop 2000™ spectrophotometer produced different readings of DNA concentration. The Qubit™ consistently estimated lower concentrations compared to the NanoDrop 2000™. Based on Qubit™ fluorometer quantification, the CheX protocol produced the highest DNA concentration (0.98 ng/μL), followed by Schi (0.962 ng/μL) and EznF (0.4325 ng/μL) (Fig. 1). The remaining protocols produced DNA concentrations < 0.3 ng/μL, with WizP and CTAB at undetectable levels. Concentrations assessed with the NanoDrop 2000™

spectrophotometer followed a similar pattern, with CheX, Schi and EznF showing the highest median DNA concentrations of 11.5 ng/μL, 6.4 ng/μL and 4.85 ng/μL, respectively. The remaining protocols had readings higher than their Qubit™ counterparts, ranging between 0.45–4.1 ng/μL (Fig. 2).

A Kruskal-Wallis non-parametric test indicated significant differences in DNA concentration between the methods investigated for NanoDrop 2000™ ($\chi^2 = 55.821$, $P = 2.218e-08$) and Qubit™ ($\chi^2 = 36.148$, $P = 1.65e-05$). A post-hoc Dunn's Multiple Comparison Test was conducted to determine significant pairwise differences between methods. Significant differences were found between several methods in NanoDrop 2000™, and

Table 2 Median DNA yield (ng/μL), total DNA yield (ng), and quality (A260/280 and A260/230) of 11 extraction protocols tested on 6 individual adult, female *Teladorsagia circumcincta* specimens

Extraction Method	Median (Range) Invitrogen Qubit™ DNA Concentration (ng/μL)	Median (Range) Invitrogen Qubit™ DNA Total Yield (ng)	Median (Range) NanoDrop 2000™ DNA Concentration (ng/μL)	Median (Range) NanoDrop 2000™ DNA Total Yield (ng)	Median (Range) NanoDrop 2000™ DNA Quality (A260/280)	Median (Range) NanoDrop 2000™ DNA Quality (A260/230)
AccM	0.0897 (0.0228–0.178)	4.485 (2.12–8.9)	0.9 (0.6–1.4)	45 (30–70)	0.935 (–13.1–1.64)	0.53 (0.18–0.73)
AccW	0.08125 (0.0129–0.127)	4.0625 (0.645–6.35)	1.3 (0.9–2.8)	65 (45–140)	1.73 (–5.76–3.58)	0.77 (0.62–1.52)
CheX	0.98 (0.429–1.64)	49 (24.6–82)	11.5 (9.3–16.5)	575 (465–825)	1.99 (1.63–2.37)	0.535 (0.46–0.67)
CTAB	Undetectable	Undetectable	0.7 (–1.2–1.1)	35 (–60–55)	0.675 (–13.83–1.51)	0.395 (–1.02–0.98)
EznF	0.4325 (0.142–0.672)	21.625 (7.1–40)	4.85 (3.2–11.5)	242.5 (160–575)	2.375 (1.99–2.83)	1.305 (0.84–2.34)
IsoG	0.2035 (0.109–0.372)	10.175 (5.45–18.15)	1.9 (1.0–3.0)	95 (50–150)	1.87 (–24.56–3.27)	1.545 (0.52–3.51)
Schi	0.962 (0.58–1.51)	48.1 (29–75.7)	6.4 (4.1–8.4)	320 (205–420)	2.19 (1.68–2.77)	0.99 (0.46–2.77)
Schi-LE	0.064 (undetectable – 0.203)	3.08 (undetectable – 10.15)	1.25 (0.5–2.3)	62.5 (25–115)	2.12 (–6.04–5.39)	0.245 (0.08–0.54)
SDS	0.16 (undetectable – 0.236)	7.425 (undetectable – 11.8)	4.1 (3.3–10.1)	205 (165–505)	3.925 (–33.69–10.76)	0.05 (0.03–0.11)
WizM	0.0079 (undetectable – 0.0079)	0 (0–0.395)	0.45 (0.2–0.9)	22.5 (10–45)	0.985 (–3.65–3.78)	–0.005 (–1.4–0.42)
WizP	Undetectable	Undetectable	1.4 (0.9–1.8)	70 (45–90)	0.505 (–12.31–17.77)	0.465 (0.26–1.06)



Qubit™ DNA concentration. Exsheathment prior to DNA extraction was found to reduce the concentration of DNA extracted by both NanoDrop 2000™ and Qubit™.

DNA purity

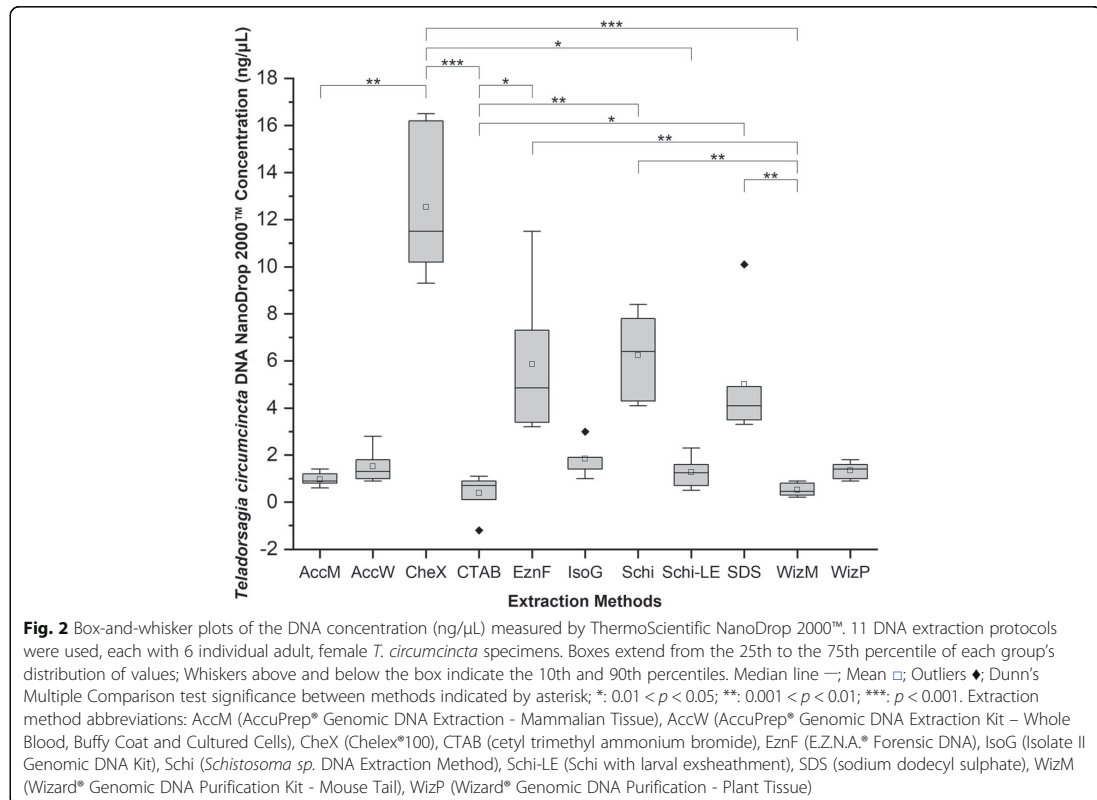
DNA purity was measured using 260 nm / 280 nm ratio (A260/280) and 260 nm / 230 nm ratio (A260/230) NanoDrop 2000™ spectrophotometer absorbency measurements. The optimal values indicating high quality DNA are 1.8 and 2.0, respectively. Most methods produced DNA with an A260/280 ranging between 1.7–2.4, with CheX, Schi and EznF methods ranging most consistently to the optimal value (Table 2, Fig. 3). IsoG was the only method whose A260/230 range met the 2.0 target value, however, the median value fell short at 1.545, indicating organic contaminants. All remaining methods ranged from – 0.005 – 1.305 (Fig. 4).

A Kruskal-Wallis non-parametric test indicated significant differences in DNA purity between the methods investigated for A260/280 ($\chi^2 = 21.102$, $P = 0.03233$), however, a post-hoc Dunn's Multiple Comparison Test determined there were no significant pairwise differences between methods in the A260/

280 measurements (Additional File 1). The Kruskal-Wallis test indicated significant differences in DNA purity between the methods investigated for A260/230 ($\chi^2 = 52.979$, $P = 1.811e-07$). Only Schi-LE, SDS and WizM were significantly different from the optimal A260/230 value (Fig. 4). Silica-binding column methods had higher quality extractions, while precipitation methods had greater DNA yield. Schi and Schi-LE indicate larval exsheathment negatively affects purity of DNA obtained as indicated by A260/280 and A260/230 analyses (Figs. 3, 4).

PCR amplification of isolated DNA

PCR success was indicated by the presence of a visible band at the expected target size of 219 bp for each sample on a 1% agarose electrophoresis gel (Fig. 5), corresponding to the *T. circumcincta* ITS-2 region. Overall, 41/66 (62.1%) of the DNA extractions were positive for *T. circumcincta* ITS-2 DNA. Six methods were consistently able to extract *T. circumcincta* DNA, with AccM, EznF, IsoG and Schi as the most consistent (6/6 samples), and AccW and CheX slightly less (5/6). Schi-LE and WizM were able to extract *T. circumcincta* DNA



half of the time, while SDS and WizP were not able to extract *T. circumcincta* DNA (Table 3). The chelating method was able to extract *T. circumcincta* DNA most of the time (5/6). Three of the silica binding column methods were 100% successful at PCR amplicon amplification, and 1 was 83.3% successful. Precipitation methods were variable, 1 was 100% successful (Schi) while others were successful in 50% of samples (Schi-LE and WizM) or less (CTAB, SDS and WizP) (Table 3). Larval exsheathment greatly reduced the reliability of DNA extraction. Four samples (CTAB 3, and WizM 1, 2, 6) which had undetectable DNA concentration when measured on Qubit™ fluorometer were positive for ITS-2 amplification (Fig. 5c).

All samples showing a positive amplification on the PCR agarose gel were sequenced. Sequencing of the approximately 219 bp fragment confirmed that 14/41 PCR products were from our target organism. 27/41 obtained fragments produced overlays (multiple peaks) and degradation of DNA was detected. 10/27 obtained fragments were below the Q20 quality cut-off. The 14 sequences which passed QC are available at GenBank (accession numbers MW161470–MW161483).

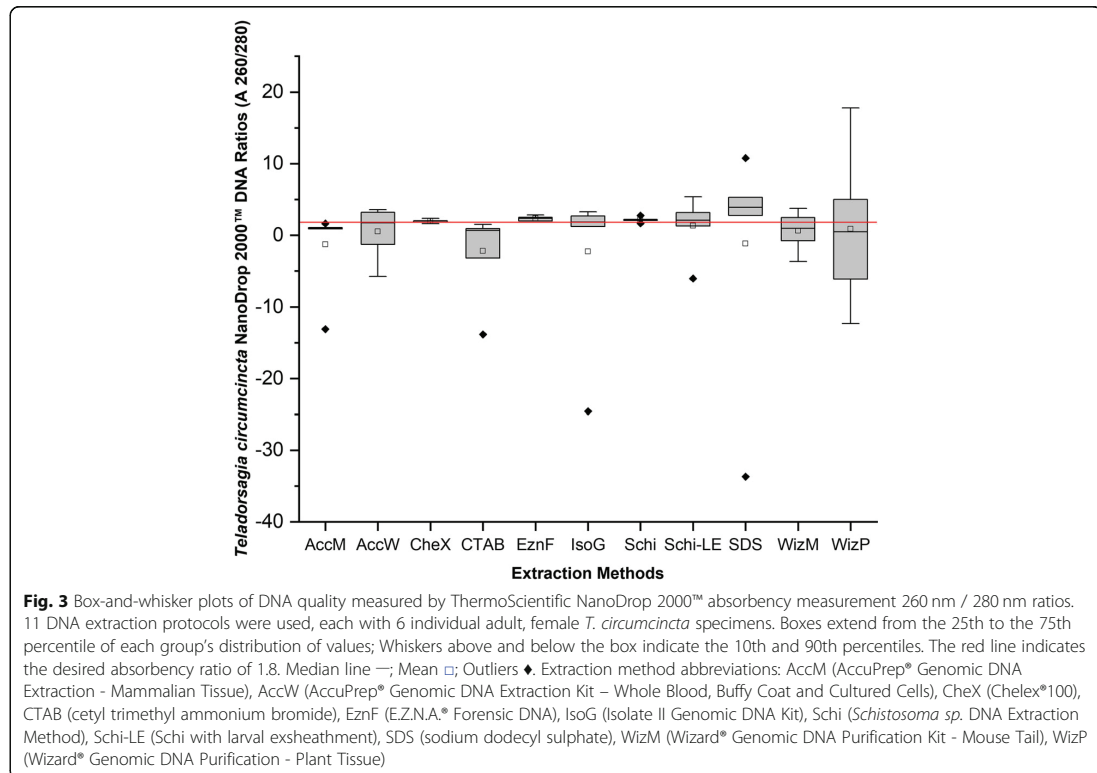
Sequence analysis

Fourteen ITS-2 sequences (Tc 1–14) from *T. circumcincta* were obtained from 9 of the DNA extraction protocols. AccM, AccW, CheX, CTAB, Schi, and WizM produced 1 sequence each, Schi-LE produced 2 sequences, and EznF and IsoG each produced 3 sequences.

Fragment lengths obtained for Tc 1–14 ranged from 53 to 186 bp. The sequences were mostly identical when aligned against the positive control reference (Fig. 6) and included segments of both ITS-2 and 28S ribosomal RNA genes. Small differences between the sequences were observed; AccM1, AccW4, CheX6, CTAB3, EznF4 and 5, IsoG3 and 6, Schi-LE2, and Schi6 substituted in base G at position 54. Additionally, WizM6 had one missing base (T) at position 122, and IsoG4 and Schi-LE1 had four additional bases (T, C, C, G) at positions 196, 199, 201 and 203, respectively. Whether these differences were due to variation in the gene or sequencing error was not determined.

Discussion

This is the first systematic comparison of different DNA extraction methods for individual *T. circumcincta* nematodes. The three types of DNA extraction tested in this



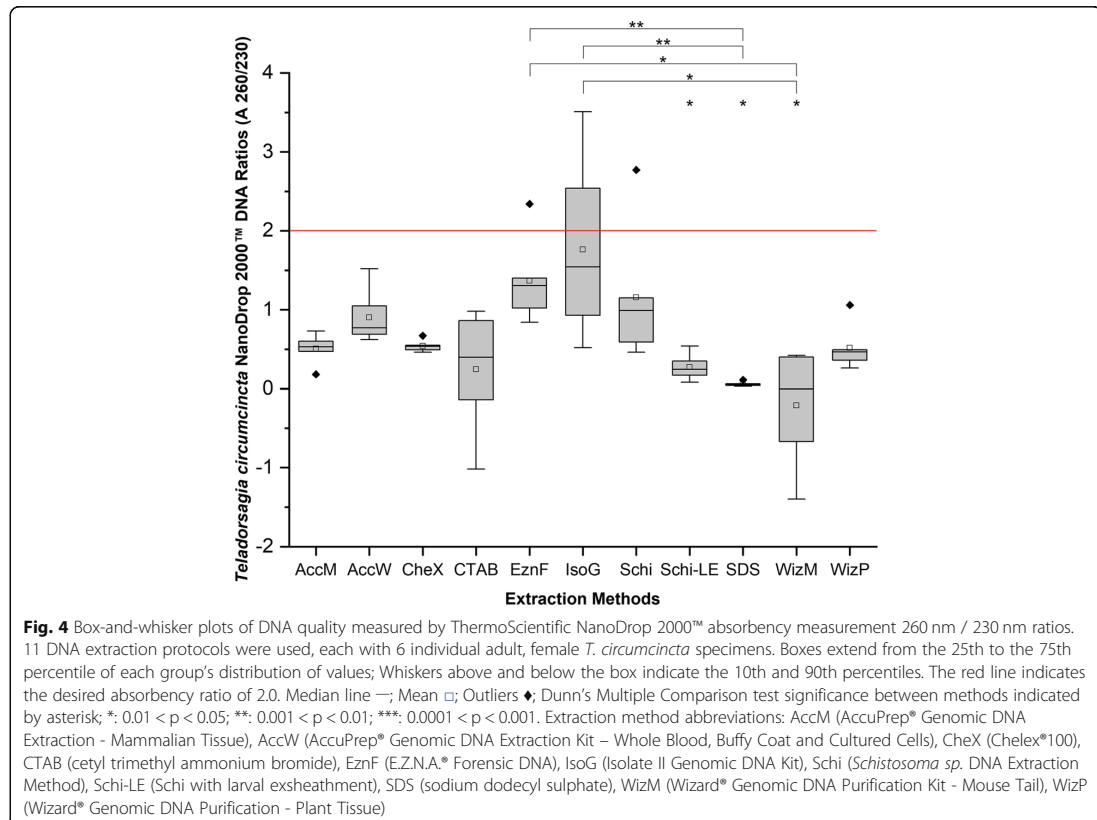
study were chelating, precipitation, and silica binding. DNA extraction method comparisons have previously been conducted on individual nematodes of other species [16], however, they did not include the traditional phenol-chloroform/CTAB method, or a chelating method as has been done in this study. The variety of DNA extraction methods available highlights the difficulty in comparing and standardizing methods between studies and organisms.

Comparison of DNA extraction methods is necessary to determine which is ideal for *T. circumcincta*. Overall, in this study silica binding column extraction methods had greater quality of *T. circumcincta* DNA, while precipitation methods were superior in terms of total DNA yield. Precipitation methods have been used on individual *T. circumcincta* specimens [27, 28], and the current reference genome for *T. circumcincta* [15]. However, other studies on *T. circumcincta*, whether pooled samples or individual specimens, have used silica binding columns [29–31]. There is currently no consensus on which DNA extraction method is best for *T. circumcincta*.

Of the 11 extraction methods compared in this study, four had relatively higher DNA concentration and

purity; CheX, EznF, IsoG and Schi, which represent all three extraction types. The processing time for these methods was on the longer end of the spectrum but were relatively short compared to methods used in other studies which have exceeded 12 h [15, 31]. The length of time needed to complete each method could be a factor determining method choice. The longer the method takes, and more hands-on the steps, determines the feasibility of a laboratory to carry out the method, and how many samples can be processed at a time or in a day. CheX was the shortest method of all, barely exceeding 30 min from start to finish. EznF, IsoG and Schi took 1.5–3 h to complete depending on the method and were more manually intensive. Schi and CheX were consistently superior in terms of DNA quantity and quality, and processing time.

Comparing quantitative and qualitative data is valuable in understanding the DNA extraction output. NanoDrop 2000™ and Qubit™ have their limitations but when used together can be powerful tools. Previous studies have claimed that spectrophotometry is the better measurement for DNA quantification [32], while most favour fluorometry [19, 33–35].



Qubit™ is based on fluorometric analysis. A fluorescent dye binds specifically to the nucleic acids within a sample and the DNA is quantified by the fluorescence measured by the detector. Even if the sample is contaminated, it can give an accurate reading because the dye is bound only to DNA. Qubit™ is highly regarded for use in sequencing and PCR because it can quantify DNA as low as 10 pg – 200 ng. The Qubit™ is considered a very accurate quantification method, more so than the NanoDrop 2000™, however, an additional instrument is required to measure the quality of a sample [36]. This study used the NanoDrop 2000™ for this additional qualitative analysis.

NanoDrop 2000™ analyses spectrophotometric absorbance. DNA absorbs light at 260 nm, however, it does not distinguish between double- and single-stranded DNA, RNA, and nucleotides. Furthermore, impurities such as protein, phenol and other salts may also measure readings at this wavelength. To account for this, the purity of DNA relative to contaminants can be determined by measuring the ratio of different wavelengths, for example, A260/280 and A260/230 [37]. The A260/280

ratio is used to determine the presence of protein in a sample, and a pure DNA A260/280 ratio should be 1.8. Lower ratios indicate protein contamination. The A260/230 ratio is used to indicate the presence of organic contaminants which could affect downstream applications. A pure DNA sample should have an A260/230 ratio of 2.0. These two ratios are used to determine the purity of a DNA sample. NanoDrop 2000™ is not as sensitive as Qubit™, quantifying 10 ng – 10 µg [37]. Additionally, calibration of the instruments is crucial to ensure the readings are as accurate as possible.

CheX, Schi and EznF had a tight IQR over the optimal A260/280 value of 1.8 indicating few contaminating proteins. However, negative A260/280 ratios were determined for several samples (Table 2). The NanoDrop 2000™ was blanked with the elution buffer of the respective method, and a negative ratio would mean that one of the wavelengths is absorbing less than zero light. This is indicative of insoluble contamination on the NanoDrop 2000™ lens. The lens was cleaned thoroughly between readings, so it could be that an insoluble contaminant has eluted with the DNA sample and caused

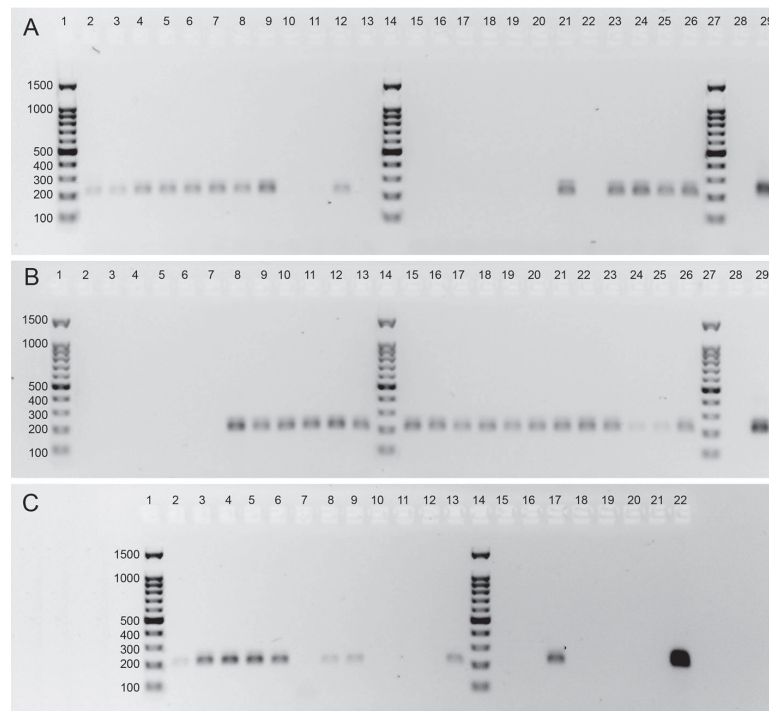
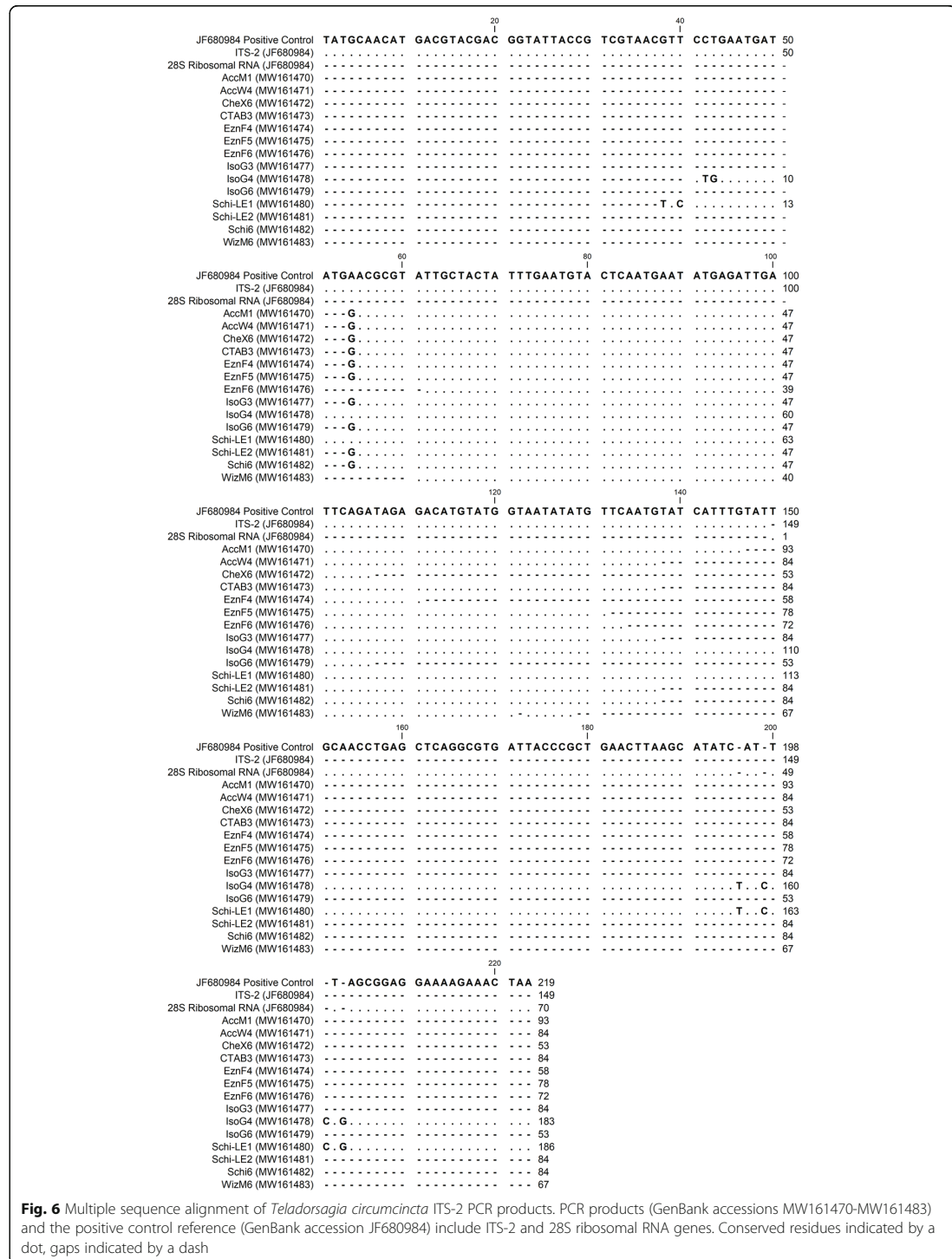


Fig. 5 Agarose gels displaying amplicons produced by conventional PCR using primer pair ITS-2/NC2. Genomic DNA samples from single *Teladorsagia circumcincta* nematode specimens. Ladder indicating amplicon size (bp) (lanes 1, 14, 27). **a** Schi (lanes 2–7), Schi-LE (lanes 8–13), SDS (lanes 15–20), CheX (lanes 21–26), no-DNA control (lane 28), *T. circumcincta* ITS-2 gBlocks™ Gene Fragment (Integrated DNA Technologies) control (lane 29). **b** WizP (lanes 2–7), EznF (lanes 8–13), AccM (lanes 15–20), IsoG (lanes 21–26), no-DNA control (lane 28), *T. circumcincta* ITS-2 gBlocks™ Gene Fragment control (lane 29). **c** AccW (lanes 2–7), WizM (lanes 8–13), CTAB (lanes 15–20), no-DNA control (lane 21), *T. circumcincta* ITS-2 gBlocks™ Gene Fragment control (lane 22). The specificity of individual amplicons produced was verified by direct sequencing

Table 3 Successful amplicon, sequencing, and median amplicon length of 11 extraction protocols tested on 6 individual adult, female *Teladorsagia circumcincta* specimens

Method	Successful Amplicon	Successful Amplicon Sequencing	Median Sequence Length (bp) (Range)
AccM	6/6 (100%)	1/6 (16.7%)	93
AccW	5/6 (83.3%)	1/5 (20%)	84
CheX	5/6 (83.3%)	1/5 (20%)	53
CTAB	1/6 (16.7%)	1/1 (100%)	84
EznF	6/6 (100%)	3/6 (50%)	72 (58–78)
IsoG	6/6 (100%)	3/6 (50%)	84 (53–183)
Schi	6/6 (100%)	1/6 (16.7%)	84
Schi-LE	3/6 (50%)	2/3 (66.7%)	135 (84–186)
SDS	0/6 (0%)	–	–
WizM	3/6 (50%)	1/3 (33.3%)	67
WizP	0/6 (0%)	–	–



inaccurate readings, or it is user error. The negative A260/280 readings were not exclusive to a method type, it was observed in both silica binding and precipitation methods. 3/4 methods of choice (CheX, EznF, and Schi) did not result in any negative A260/280 readings, further confirming their superiority amongst the methods compared in this study whether because the method does not allow, or user error does not result in, insoluble contaminants. The decision to retain these samples and results, and not do a new round of extractions was to show what results are possible with each of these methods; the good, the bad and the ugly.

The A260/230 measurements found IsoG as the only extraction method whose IQR encompassed the optimal value of 2.0, however, this IQR was wide and the median low at 1.545, so although some individual samples may be acceptable, it is overall a poor result. All methods struggled to minimize organic contaminants, and this can greatly impact downstream applications. A260/230 is affected by the salinity of the elution buffer. Increased salt concentration in the DNA sample will lower the A260/230 because of salt absorbance at 230 nm [38]. Both precipitation and silica-binding methods use salts to precipitate and bind DNA, respectively [22]. Additional wash steps may be required to remove excess and contaminating salts.

Traditionally, DNA purity has been prioritised over concentration because of the effects of contamination on downstream applications [39, 40], and prioritising concentration over purity may result in an inaccurate concentration reading due to contaminants. Additional, gentle, washing steps may increase the purity of the sample, if it is required, but there is a risk of decreasing DNA concentration. For samples which contain additional DNA from non-target organisms, concentration may become more of a priority. RL Smith et al. [19] found that because their target ancient powdery mildew DNA was a tiny proportion of the total DNA extracted from the plant specimen, a higher total DNA concentration increased the chance of sequencing powdery mildew DNA. In this study, extracting DNA from individual nematode specimens used the entire worm, additional material cannot be added to improve the DNA concentration and there is likely to be bacteria present due to *T. circumcincta* being collected from the host gut. The methods most successful at PCR amplification were silica binding, and we found these methods had overall higher quality than precipitation or chelating, as has been seen in other studies [17, 18]. NanoDrop 2000™ likely overestimated the concentration of DNA in the samples and the Qubit™ concentration of *T. circumcincta* DNA extracted was low irrespective of the extraction method chosen. As long as purity of a sample is reasonably acceptable, we recommend prioritising a higher

DNA concentration over achieving total purity to ensure as much *T. circumcincta* DNA is being extracted as possible because of the presence of host or microorganism DNA that is likely to be present. Calculating correct DNA concentration and purity when concentrations are low is difficult and unlikely to be completely accurate. Additionally, DNA concentration should be based on Qubit™ readings as they are more likely to be a true representation of DNA in a sample than NanoDrop 2000™ readings.

PCR amplification of the ITS-2 gene provides confirmation of correct species DNA extraction, as well as indicating PCR capability of the extraction. For example, impurities such as protein, phenol and other salt traces may terminate a PCR reaction. The methods which had high concentration and purity had the best PCR success, and most methods with low concentration and purity were not PCR successful. There does not appear to be any correlation that indicates a higher concentration or purity of DNA was more successful at PCR amplification. Schi, EznF, AccM and IsoG were all 100% successful at PCR ITS-2 amplification (Fig. 5). Schi, EznF and IsoG PCR success is not unexpected given their generally higher quality DNA. AccM, on the other hand, had low A260/280 (Fig. 3), and very low A260/230 (Fig. 4), but the quality did not seem to affect PCR capability. Interestingly, CTAB had the second lowest concentration on NanoDrop 2000™, was undetectable on Qubit™ and was very low in purity for both A260/280 and A260/230 (Figs. 3, 4), but was still able to produce one successful ITS-2 amplicon for *T. circumcincta* (Tables 2, 3, Fig. 5), indicating that despite the odds, this method can extract *T. circumcincta* DNA, though unreliably. Sequencing of the PCR products had an overall poor outcome; few were sequenced successfully, and the successful fragments were far shorter than expected. Of the four methods of interest, IsoG and Schi shared the greatest median sequence length at 84 bp, much shorter than the target amplicon length of 219 bp. It is possible shearing of the sequences occurred. AccM and Schi extractions produced a sequence of high enough quality only 16.7% of the time. Whereas EznF and IsoG produced sequences of high enough quality 50% of the time. It is unclear why 27 of the PCR products did not meet quality standards, however, it is likely due to the PCR clean-up kit and method used. Initially the PCR products were eluted into TE buffer and stored. When prepared for sequencing, the samples were washed in ethanol and eluted into nuclease-free water. Either the DNA clean-up method, the temporary storage in TE buffer, or the additional wash step has contributed to the background interference and low quality of sequences reported by AGRF.

There are a range of factors here that are influencing sequencing success, and the methods and techniques used in this study have room for optimisation and improvement.

Conclusions

This study highlights the difficulties in extracting DNA from *T. circumcincta* and that the different extraction methods tested vary significantly in the quality and quantity of DNA recovered. We found the Schi method for extracting DNA from individual *T. circumcincta* nematode specimens to be the best. Schi was able to extract a relatively high concentration of DNA when measured on both NanoDrop 2000™ and Qubit™ (Figs. 1, 2), and measured a tight IQR over the optimal A260/280 (Fig. 3). The CheX method was a strong contender and is a close second, however, it was not as successful at ITS-2 PCR amplification as Schi (Table 3). Considering the purpose is to reliably extract *T. circumcincta* DNA, successful ITS-2 amplification is an important contributing factor. Additionally, the Schi method is simple and fast at approximately 2 h. Exsheathment of the *T. circumcincta* cuticle is unnecessary. Although not tested in this study, it is likely these results could be generalized to related species such as *Haemonchus contortus* or *Ostertagia ostertagi*. Further optimization of methods for extracting *T. circumcincta* DNA is required.

Methods

Sampling

Teladorsagia circumcincta specimens were harvested from experimentally infected sheep at Federation University, Australia). Animals were bred and supplied on-site by the Gippsland Field Station, owned and operated by Monash University. Animal ethics approval was approved by Federation University (AEC # 17008). 5–6-month-old Merino wethers were infected with 5000 *T. circumcincta* third-stage larvae. Approximately 5 weeks post infection, adult parasites were collected from the abomasum. The animals were euthanized (humanely killed) with a lethal intravenous dose (5 g) of pentobarbitone (Lethobarb®, Virbac Pty Ltd., Sydney, Australia) administered by jugular venepuncture. This is a standard method to induce euthanasia in animals. The abomasum was removed from the animal and opened along the greater curvature to reveal the gastric mucosa. The abomasum was then placed in a 50 × 30 cm plastic tray and gently rinsed with saline to remove the contents. All folds of the abomasum were carefully examined to collect all parasites. Parasites dislodged from the abomasal surface following washing were collected with forceps. All collected parasites were washed in PBS 3 times at

4 °C. *T. circumcincta* specimens were separated by sex by light microscopy and stored in 100% ethanol at 4 °C and washed in MilliQ H₂O immediately prior to use.

DNA extraction

Ten DNA extraction protocols were selected for comparison, which encompassed the most common DNA extraction methods, including chelating, silica binding and precipitation (Table 1). An exsheathment step prior to DNA extraction was applied to 1 DNA extraction protocol to determine whether cuticle removal improved DNA yield, bringing the total protocols compared to 11. The methods were performed as per both manufacturer's instructions and previously published DNA extraction protocols. Six individual adult, female nematode specimens were used for each extraction method. Males were excluded for a separate study. Detailed methods are outlined in Additional File 2. All methods were eluted to 50 µL. DNA was quantified using two methods, Qubit™ fluorometer (Life Technologies, Singapore) and NanoDrop 2000™ spectrophotometer (Thermo Fisher Scientific, Wilmington, Delaware, USA). The NanoDrop 2000™ was blanked using the respective elution buffer for the method. All DNA extraction product concentrations were measured using Qubit™ (1X dsDNA HS (High Sensitivity) Assay Kit, Invitrogen, #Q33231) and NanoDrop 2000™, and purity was measured using the 260/280 nm and 260/230 nm absorbency ratios of NanoDrop 2000™.

Statistical analysis

Statistical analyses were performed using R version 4.03 [41] and package FSA [42]. The distribution of the data was determined to be non-normally distributed using the graphics package and function 'hist'. The function 'kruskal.test' was used to perform a Kruskal-Wallis non-parametric test to identify significant differences among DNA extraction methods in DNA concentration when measured on NanoDrop 2000™ spectrophotometer and Qubit™ fluorometer, 260/280 nm absorbance ratio, and 260/230 nm absorbance ratio. The function 'dunn.test' was used to perform a post-hoc Dunn's Multiple Comparison Test to identify significant pairwise comparisons. Variables were considered significant with $P \leq 0.05$. P -values adjusted with the Holm method.

PCR amplification and sequencing of the rRNA ITS-2 from individual nematodes

An approximately 219 bp expected fragment size encompassing partial ITS-2 and 28S ribosomal RNA sequences was amplified by PCR from individual nematodes using the ITS-2 (5'-TATGCAACATGACGTACGACGG-3') and NC2 (5'-TTAGTTTCTTTTCTCCGC-3') primers described in NJ Bott et al. [31]. PCR reaction conditions

were 12.5 µL GoTaq Green Master Mix (Promega, M7122), 0.5 µL ITS-2 primer (10 µM), 0.5 µL NC2 primer (10 µM), 2.5 µL DNA template, and 9 µL nuclease-free H₂O. The negative control contained nuclease-free H₂O, and the positive control contained the following: a gBlocks™ Gene Fragment (Integrated DNA Technologies, Coralville, USA) of the last 219 bp of *Teladorsagia circumcincta* 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence (GenBank accession: JF680984) in place of the DNA template, respectively. The thermocycling parameters were 95 °C for 2 min, followed by 50 cycles of 95 °C for 15 s, 60 °C for 30 s, 72 °C for 15 s followed by a final extension of 72 °C for 5 min. PCR products were confirmed on 1% agarose gel. Positive PCR products were purified with a Wizard® SV Gel and PCR Clean-Up System (Promega, A9281) and temporarily stored in TE buffer at 4 °C. PCR products were washed in ethanol and eluted into nuclease-free water prior to being sent to the Australian Genome Research Facility and directly sequenced using Sanger sequencing with the ITS-2 primer. Sequences were aligned using Geneious version 2020.2 created by Biomatters. Available from <https://www.geneious.com>.

Sequenced PCR products which passed QC (≥Q20) were aligned with the PCR positive control using Geneious Prime software.

Abbreviations

DNA: Deoxyribonucleic acid; PCR: Polymerase chain reaction; AccM: AccuPrep® Genomic DNA Extraction - Mammalian Tissue; AccW: AccuPrep® Genomic DNA Extraction Kit - Whole Blood, Buffy Coat and Cultured Cells; CheX: Chelex®100; CTAB: Cetyl trimethyl ammonium bromide; Eznf: E.Z.N.A.® Forensic DNA; IsoG: Isolate II Genomic DNA Kit; Schi: *Schistosoma* sp. DNA Extraction Method; Schi-LE: Schi with larval exsheathment; SDS: sodium dodecyl sulphate; WizM: Wizard® Genomic DNA Purification Kit - Mouse Tail; WizP: Wizard® Genomic DNA Purification - Plant Tissue; ITS-2: Internal transcribed spacer region 2; A260/280: 260 nm / 280 nm absorbency ratio; A260/230: 260 nm / 230 nm absorbency ratio; Q20: Q-score ≥20; QC: Quality check; T: Thymine; C: Cytosine; G: Guanine; TE buffer: Tris-EDTA buffer; AGRF: Australian Genome Research Facility; PBS: Phosphate-buffered saline; rRNA: Ribosomal ribonucleic acid

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12896-021-00695-6>.

Additional file 1: Table S1. Statistically significant differences in NanoDrop 2000™ and Qubit™ DNA concentration, and 260/230 nm absorbance ratio using the 11 DNA extraction methods according to the Dunn's Multiple Comparison Test.

Additional file 2. DNA extraction methods.

Acknowledgements

An anonymous reviewer is thanked for critically reading the manuscript and suggesting substantial improvements.

Authors' contributions

SS, CJ, and MS conceived the study. SS designed and performed the research and data analysis, and wrote the manuscript. DP and SP undertook the sampling procedures. CJ and MS verified the analytical methods and supervised the findings of this work. All authors read and approved the final manuscript.

Funding

This research was funded by a start-up grant from La Trobe University. The University played no role in the design of the study or in the collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

All data generated or analyzed during this study are included in this article and its supplementary information files (Additional files 1, 2) or are available in the GenBank repository (GenBank accessions MW161470-MW161483). Raw datasets used and/or analyzed during the current study are available from the corresponding author.

Declarations

Ethics approval and consent to participate

All procedures involving animals were approved by the Animal Ethics Committee of Federation University (17–008).

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

¹AgriBio Centre for AgriBioscience, Department of Animal, Plant and Soil Sciences, School of Life Sciences, La Trobe University, 5 Ring Road, Bundoora, Victoria 3086, Australia. ²School of Science, Psychology and Sport, Federation University, Churchill, Victoria, Australia.

Received: 29 December 2020 Accepted: 9 May 2021

Published online: 17 May 2021

References

- Lane J, Jubb T, Shephard R, Webb-Ware J, Fordyce G. GHD Pty Ltd: final report: priority list of endemic diseases for the red meat industries. North Sydney: Meat & Livestock Australia Limited; 2015.
- Jackson F, Coop RL. The development of anthelmintic resistance in sheep nematodes. *Parasitology*. 2000;120(Suppl):S95–107.
- Bartley DJ, Jackson E, Johnston K, Coop RL, Mitchell GB, Sales J, et al. A survey of anthelmintic resistant nematode parasites in Scottish sheep flocks. *Vet Parasitol*. 2003;117(1–2):61–71. <https://doi.org/10.1016/j.vetpar.2003.07.023>.
- Leathwick DM, Besier RB. The management of anthelmintic resistance in grazing ruminants in Australasia—strategies and experiences. *Vet Parasitol*. 2014;204(1–2):44–54. <https://doi.org/10.1016/j.vetpar.2013.12.022>.
- 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(3):479–88. [https://doi.org/10.1016/0021-9975\(83\)90035-X](https://doi.org/10.1016/0021-9975(83)90035-X).
- 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(2):235–45. [https://doi.org/10.1016/0021-9975\(85\)90010-6](https://doi.org/10.1016/0021-9975(85)90010-6).
- Smith WD, Jackson F, Jackson E, Graham R, Williams J, Willadsen SM, 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(3):300–6. [https://doi.org/10.1016/S0034-5288\(18\)30620-9](https://doi.org/10.1016/S0034-5288(18)30620-9).
- Stear MJ, Bairden K, Bishop SC, Buitkamp J, Duncan JL, Gettinby G, et al. The genetic basis of resistance to *Ostertagia circumcincta* in lambs. *Vet J*. 1997;154(2):111–9. [https://doi.org/10.1016/S1090-0233\(97\)80049-4](https://doi.org/10.1016/S1090-0233(97)80049-4).
- Stear MJ, Doligalska M, Donskow-Schmelter K. Alternatives to anthelmintics for the control of nematodes in livestock. *Parasitology*. 2007;134(Pt 2):139–51. <https://doi.org/10.1017/S0031182006001557>.

10. 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(1):45–52. <https://doi.org/10.1079/AHRR200351>.
11. Taylor MA, Coop RL, Wall RL. *Veterinary parasitology*, 3rd edn. Oxford, UK: Ames: Blackwell; 2007.
12. Stear MJ, Bishop SC. The curvilinear relationship between worm length and fecundity of *Teladorsagia circumcincta*. *Int J Parasitol*. 1999;29(5):777–80. [https://doi.org/10.1016/S0020-7519\(99\)00019-3](https://doi.org/10.1016/S0020-7519(99)00019-3).
13. Parkinson J, Mitreva M, Whitton C, Thomson M, Daub J, Martin J, et al. A transcriptomic analysis of the phylum Nematoda. *Nat Genet*. 2004;36(12):1259–67. <https://doi.org/10.1038/ng1472>.
14. Palevich N, Maclean PH, Mitreva M, Scott R, Leathwick D. The complete mitochondrial genome of the New Zealand parasitic roundworm *Teladorsagia circumcincta* (Trichostrongyloidea: Haemonchidae) field strain NZ_Teci_NP. *Mitochondrial DNA B Resour*. 2019;4(2):2869–71. <https://doi.org/10.1080/23802359.2019.1660241>.
15. Choi YJ, Bisset SA, Doyle SR, Hallsworth-Pepin K, Martin J, Grant WN, et al. Genomic introgression mapping of field-derived multiple-anthelmintic resistance in *Teladorsagia circumcincta*. *PLoS Genet*. 2017;13(6):e1006857. <https://doi.org/10.1371/journal.pgen.1006857>.
16. Doyle SR, Sankaranarayanan G, Allan F, Berger D, Jimenez Castro PD, Collins JB, et al. Evaluation of DNA extraction methods on individual Helminth egg and larval stages for whole-genome sequencing. *Front Genet*. 2019;10:826. <https://doi.org/10.3389/fgene.2019.00826>.
17. Seesao Y, Audebert C, Verrez-Bagnis V, Merlin S, Jerome M, Viscogliosi E, et al. Monitoring of four DNA extraction methods upstream of high-throughput sequencing of Anisakidae nematodes. *J Microbiol Methods*. 2014;102:69–72. <https://doi.org/10.1016/j.jmimet.2014.05.004>.
18. Schiebelhut LM, Abboud SS, Gomez Daglio LE, Swift HF, Dawson MN. A comparison of DNA extraction methods for high-throughput DNA analyses. *Mol Ecol Resour*. 2017;17(4):721–9. <https://doi.org/10.1111/1755-0998.12620>.
19. Smith RL, Sawbridge T, Mann R, Kaur J, May TW, Edwards J. Rediscovering an old foe: optimised molecular methods for DNA extraction and sequencing applications for fungarium specimens of powdery mildew (Erysiphales). *PLoS One*. 2020;15(5):e0232535. <https://doi.org/10.1371/journal.pone.0232535>.
20. Walsh PS, Metzger DA, Higuchi R. Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *Biotechniques*. 1991;10(4):506–13.
21. Lienhard A, Schaffer S. Extracting the invisible: obtaining high quality DNA is a challenging task in small arthropods. *PeerJ*. 2019;7:e6753. <https://doi.org/10.7717/peerj.6753>.
22. Karp A, Isaac PG, Ingram DS. Isolation of Nucleic Acids Using Silica-Gel Based Membranes: Methods Based on the Use of QIAamp Spin Columns. In: Karp A, Isaac PG, Ingram DS, editors. *Molecular Tools for Screening Biodiversity*. Dordrecht: Springer Netherlands; 1998. p. 59–63.
23. Sarkinen T, Staats M, Richardson JE, Cowan RS, Bakker FT. How to open the treasure chest? Optimising DNA extraction from herbarium specimens. *PLoS One*. 2012;7(8):e43808. <https://doi.org/10.1371/journal.pone.0043808>.
24. Brindley PJ, Lewis FA, McCutchan TF, Bueding E, Sher A. A genomic change associated with the development of resistance to hycanthone in *Schistosoma mansoni*. *Mol Biochem Parasitol*. 1989;36(3):243–52. [https://doi.org/10.1016/0166-6851\(89\)90172-2](https://doi.org/10.1016/0166-6851(89)90172-2).
25. Edwards K, Johnstone C, Thompson C. A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. *Nucleic Acids Res*. 1991;19(6):1349. <https://doi.org/10.1093/nar/19.6.1349>.
26. Dawkins HJ, Spencer TL. The isolation of nucleic acid from nematodes requires an understanding of the parasite and its cuticular structure. *Parasitol Today*. 1989;5(3):73–6. [https://doi.org/10.1016/0169-4758\(89\)90005-7](https://doi.org/10.1016/0169-4758(89)90005-7).
27. Gasser RB, Chilton NB, Hoste H, Beveridge I. Rapid sequencing of rDNA from single worms and eggs of parasitic helminths. *Nucleic Acids Res*. 1993;21(10):2525–6. <https://doi.org/10.1093/nar/21.10.2525>.
28. Stevenson LA, Gasser RB, Chilton NB. The ITS-2 rDNA of *Teladorsagia circumcincta*, *T. trifurcata* and *T. davitiani* (Nematoda: Trichostrongylidae) indicates that these taxa are one species. *Int J Parasitol*. 1996;26(10):1123–6. [https://doi.org/10.1016/S0020-7519\(96\)80013-0](https://doi.org/10.1016/S0020-7519(96)80013-0).
29. Martinez-Valladares M, Valderas-Garcia E, Gandasegui J, Skuce P, Morrison A, Castilla Gomez de Agüero V, Cambra-Pelleja M, Balana-Fouce R, Rojo-Vazquez FA: *Teladorsagia circumcincta* beta tubulin: the presence of the E198L polymorphism on its own is associated with benzimidazole resistance. *Parasit Vectors*. 2020;13(1):453. <https://doi.org/10.1186/s13071-020-04320-x>.
30. Ashrafi K, Sharifdini M, Heidari Z, Rahmati B, Kia EB. Zoonotic transmission of *Teladorsagia circumcincta* and *Trichostrongylus* species in Guilan province, northern Iran: molecular and morphological characterizations. *BMC Infect Dis*. 2020;20(1):28. <https://doi.org/10.1186/s12879-020-4762-0>.
31. Bott NJ, Campbell BE, Beveridge I, Chilton NB, Rees D, Hunt PW, et al. A combined microscopic-molecular method for the diagnosis of strongylid infections in sheep. *Int J Parasitol*. 2009;39(11):1277–87. <https://doi.org/10.1016/j.ijpara.2009.03.002>.
32. Haque KA, Pfeiffer RM, Beerman MB, Struewing JP, Chanock SJ, Bergen AW. Performance of high-throughput DNA quantification methods. *BMC Biotechnol*. 2003;3(1):20. <https://doi.org/10.1186/1472-6750-3-20>.
33. O'Neill M, McPartlin J, Arthure K, Riedel S, McMillan ND. Comparison of the TLDA with the Nanodrop and the reference Qubit system. *J Phys Conf Ser*. 2011;307:012047. <https://doi.org/10.1088/1742-6596/307/1/012047>.
34. Vilkkij J, Uimari P, Sironen A. Comparison of different DNA extraction methods from hair root follicles to genotype Finnish landrace boars with the Illumina PorcineSNP60 BeadChip. *Agric Food Sci*. 2008;20(2):143–50.
35. Nakayama Y, Yamaguchi H, Einaga N, Esumi M. Pitfalls of DNA quantification using DNA-binding fluorescent dyes and suggested solutions. *PLoS One*. 2016;11(3):e0150528. <https://doi.org/10.1371/journal.pone.0150528>.
36. Simbolo M, Gottardi M, Corbo V, Fasan M, Mafficini A, Malpeli G, et al. DNA quantification workflow for next generation sequencing of histopathological samples. *PLoS One*. 2013;8(6):e62692. <https://doi.org/10.1371/journal.pone.0062692>.
37. Matlock B. Assessment of nucleic acid purity. Technical Note 52646. In: Thermo Fisher Scientific Inc; 2015.
38. Lucena-Aguilar G, Sanchez-Lopez AM, Barberan-Aceituno C, Carrillo-Avila JA, Lopez-Guerrero JA, Aguilar-Quesada R. DNA source selection for downstream applications based on DNA quality indicators analysis. *Biopreserv Biobank*. 2016;14(4):264–70. <https://doi.org/10.1089/bio.2015.0064>.
39. Hiesinger M, Löffert D, Ritt C, Oelmüller U. The effects of phenol on nucleic acid preparation and downstream applications. *Qiagen News*. 2001;5:23–6.
40. Mahmoudi N, Slater GF, Fulthorpe RR. Comparison of commercial DNA extraction kits for isolation and purification of bacterial and eukaryotic DNA from PAH-contaminated soils. *Can J Microbiol*. 2011;57(8):623–8. <https://doi.org/10.1139/w11-049>.
41. R Development Core team: R: a language and environment for statistical computing. In: Vienna, Austria: R Foundation for statistical Computing; 2020.
42. Ogle DH, Wheeler P, Dinno A: FSA: Fisheries Stock Analysis <https://github.com/droglenc/FSA>: R package version 0.8.31; 2020.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Chapter 4

Teladorsagia circumcincta genome sequencing on the MinION™ and microbiome
species identification

4.1 Introduction

Teladorsagia circumcincta is the most important parasitic nematode of sheep in cool temperate regions worldwide (Bartley *et al.*, 2003). *T. circumcincta* infection results in reduced production, impaired animal welfare, parasitic gastroenteritis, poor growth performance, and weight loss (Stear *et al.*, 2003). To develop new control strategies for *T. circumcincta* infection, it is pivotal that research uncovers as much information about the biology of this nematode as possible, and an important starting point involves the genomic investigation of *T. circumcincta*.

T. circumcincta has a standard one-host lifecycle as explained in Chapter 1. Briefly, eggs are passed in faeces onto pasture where they develop into pre-infective larval stages and then into infective third-stage larvae (L3) which are ingested by the host. L3 exsheath in the rumen of the host and enter the lumen of the abomasal glands 2-3 days after ingestion, where they develop into the pre-adult and immature adult (L5) stages and establish within the host. L5 then mature into sexually active adults which feed and breed on the mucosal surface of the abomasa (Marchiondo *et al.*, 2019; Venturina *et al.*, 2013). Attempts at culturing *T. circumcincta* entirely *in vitro* have been unsuccessful. Studies examining the adult stages of *T. circumcincta* require short-term cultures of adults harvested from the host *in vivo* (Luque *et al.*, 2010). Consequently, producing clean nematodes for uncontaminated DNA extraction and sequencing is almost impossible.

Our previous study, Chapter 2, used the available draft genome (BioProject PRJNA72569) by Choi *et al.* (2017) to look specifically at the cathepsin F gene of *T. circumcincta* and found the draft genome to be incomplete with many gaps, at times misassembled, and scarcely annotated (Sloan *et al.*, 2020). This draft genome was assembled using short-read technology. Short read sequences generally have lower sequencing error rates (Dominguez Del Angel *et al.*, 2018; Kchouk *et al.*, 2017; Quail *et al.*, 2012), however, they often fail to generate sufficient overlap in the DNA fragments causing major problems for *de novo* assembly as discontinuous contigs or large repetitive regions may occur (Adewale, 2020). Re-assembly using data from BioProject PRJEB7676 had some success for cathepsin F, but ultimately determined more data was required to fill in the gaps (Sloan *et al.*, 2020).

The Oxford Nanopore Technology (ONT) MinION™ device utilises a nanopore for sequencing; a nano-scale hole with a size of approximately 1.4 nm diameter (Liu *et al.*, 2012). MinION™ devices contain a protein nanopore set in an electrically resistant polymer membrane or flow cell. The membrane is immersed in an electrolyte solution, and an ionic current is passed through the nanopore by setting a voltage across the membrane. Single molecules entering the nanopore cause characteristic disruptions in the current, and by measuring this disruption, DNA or RNA molecules can be identified and characterised (Kasianowicz *et al.*, 1996; Vilgis *et al.*, 2018). ONT sequencers are currently able to produce the longest reads on the market, at upwards of 2 Mbp,

and are often referred to as “ultra-long reads” (Kraft *et al.*, 2019). These long reads are useful for closing gaps that short read genome assemblies contain. ONT has not previously been used to sequence *T. circumcincta*, and this study explores this new technology, the difficulties which arise from sequencing an obligate parasite, and the additional contaminating species present.

4.2 Materials & Methods

4.2.1 Sampling

Teladorsagia circumcincta specimens were harvested from experimentally infected sheep at Federation University, Australia. Animals were bred and supplied on-site by the Gippsland Field Station, owned and operated by Monash University. Animal ethics was approved by Federation University (AEC # 17008). 5 – 6-month-old Merino wethers were infected with 5,000 *T. circumcincta* third-stage larvae. Approximately 5 weeks post infection, adult parasites were collected from the abomasum. The animals were euthanized with a lethal intravenous dose (5 g) of pentobarbitone (Lethobarb®, Virbac Pty Ltd., Sydney, Australia) administered by jugular venepuncture. The abomasum was removed from the animal and opened along the greater curvature to reveal the gastric mucosa. The abomasum was then placed in a 50 x 30 cm plastic tray and gently rinsed with saline to remove the contents. All folds of the abomasum were carefully examined to collect all parasites. Parasites dislodged from the abomasal surface following washing were collected with forceps. All collected parasites were washed in PBS 3 times at 4°C. Light microscopy was used to separate *T. circumcincta* specimens by sex, stored in 100% ethanol at 4°C and washed in MilliQ H₂O immediately prior to use.

4.2.2 DNA extraction

DNA was extracted from individual adult *T. circumcincta* specimens using a method described for isolation of DNA from *Schistosoma spp.* (Bender *et al.*, 1983; Sloan *et al.*, 2021). In brief, single male worms were homogenized in 1.5 mL Eppendorf tubes with a Cordless Pellet Pestle (Kimble Chase Life Science and Research Products LLC, Vineland, USA) in 50 µL 50 mM Tris, pH 8.0, 100 mM NaCl, 25 mM sucrose, 10 mM EDTA, 2% SDS. Following incubation at 65°C for 30 min, ammonium acetate was added to 1 M, and the tube transferred to an ice bath for 30 min. The tube was then centrifuged at 3,000 x g in an Eppendorf Centrifuge 5424 R for 10 min at 4°C to pellet precipitated proteins, and the DNA was isolated from the supernatant by ethanol precipitation. The DNA precipitate was washed with 100% ethanol, air dried, and dissolved in 20 µL of 10 mM Tris, pH 8.0, 0.1 mM EDTA (TE buffer). DNA quantity and quality was measured using a NanoDrop 2000™ spectrophotometer and Qubit™ fluorometer, and total remaining yield was carried over for sequencing.

4.2.3 Sequencing on MinION™

Three individual *T. circumcincta* DNA extractions were prepared for separate MinION™ runs: TcM1, 2, and 3, with 344 ng, 158 ng and 196 ng of template DNA, respectively. Library preparation was performed with Ligation Sequencing Kit SQK-LSK109 (Oxford Nanopore Technologies, ONT, Oxford, UK) following manufacturer's protocol.

11 individual *T. circumcincta* DNA extractions were prepared for one barcoded MinION™ run, TcM4, with a combined 1.59 µg of starting template DNA. Library preparation was performed with PCR Barcoding Kit SQK-PBK004 following manufacturer's protocol. TcM4 was not included in microbiome analysis because the PCR library preparation artificially favours specific sequences primers are targeted towards and is not a true representation of DNA coverage present in the samples.

Reads acquisition (ONT's MinKNOW core ver. 3.4.8) and base-calling (ONT's Guppy software ver. 3.3.0) was integrated on ONT's MinION™ device (MinION™ Mk1B) connected via USB3.0 to a Dell Latitude 7490 laptop (Intel i7-8650U CPU, 2.11 GHz, 16 GB RAM with 1 TB SSD storage running Windows 10 Enterprise). The laptop was used primarily to run MinKNOW. Read acquisition was terminated when sequencing state time equivalent reached ~2% at approximately 25 hours.

4.2.4 Read classification

Basecalled reads were submitted to ONT's cloud-based EPI2ME What's In My Pot? (WIMP) workflow (Juul *et al.*, 2015) and separated into classified and unclassified reads. Classified reads were grouped into Kingdom, Phylum, and Class. Reads classified as *Homo sapiens* were used as separate queries for BLASTn analysis against the NCBI nucleotide database with an e-value threshold of 1×10^{-9} .

4.2.5 Genome assembly

PacBio RS sequence reads (BioProject PRJEB7676; Leinonen *et al.*, 2011) were obtained from the Sequence Read Archive. Unclassified TcM1-4 and PacBio reads were assembled using Canu 2.1.1 (Koren *et al.*, 2017) with genome size value set at 58.6 Mb (S4.1 *Supplementary File 1*). Protein-coding genes were predicted using Augustus 3.0.2 (S4.2 *Supplementary File 2*) with *Caenorhabditis elegans* as the reference organism (Stanke *et al.*, 2003). Contigs were analysed and annotated using Geneious Prime version 2021.1.1 created by Biomatters. Available from <https://www.geneious.com>.

4.2.6 Cathepsin F gene

Genome assembly contigs were used to create a BLASTn database in Geneious Prime. Secreted cathepsin F protein sequence (Tci-CF-1, GenBank accession: ABA01328) was used as a query for pBLASTn against the contig database. A multiple sequence alignment of Tci-CF-1 and predicted

cathepsin F genes from the contig database was performed in Geneious Prime. Selected Augustus predicted genes were used as separate queries for BLASTn against the NCBI nucleotide database.

4.3 Results

4.3.1 Sequence basecalling analysis

A total of 3,687,435 reads were basecalled by the MinION™ and guppy software. Quality score cut-off was set at 7 (Q7). 729,217 reads were below Q7 and removed from further analysis. The total yield was 4.4 Gbp, with an average sequence length of 5,151 bp and average quality score of 10.71. A breakdown of each individual run is displayed in Table 4.1.

Table 4.1: Oxford Nanopore MinION™ sequencing data and What's In My Pot? read classification data.

Sequence Run		TcM 1	TcM 2	TcM 3	TcM 4
Total Reads		17,258	379,716	16,536	3,273,925
N50 (Kbp)		38.67	2.78	2.86	1.22
GC Content (%)		48.7	43.4	43.8	46.1
Reads >Q7		17,155	379,716	16,530	2,544,817
Total Yield (Gbp)		0.235	1.1	0.053	3.0
Length (bp)	Average	13,606	2,849	3,219	931
	Longest	102,678	51,851	42,848	11,318
Quality Score		10.68	9.97	12.27	9.9
Reads					
Classified	Total	12,578	231,797	10,801	517,373
	Bacteria	9,056	227,805	10,468	452,952
	Virus	2,515	1,388	48	413
	Eukaryote	881	2,598	284	59,284
	Archaea	126	6	1	187
Unclassified	Total	4,577	52,022	2,660	2,018,331

TcM1 had a total of 17,258 basecalled reads. 103 reads were below Q7 and removed from further analysis. The total yield was 234.8 Mbp, with an average sequence length of 13,606 bp, and average quality score of 10.68. The longest read was 102,678 bp in length, and the N50 was 38.67 Kbp. Guanine-cytosine (GC) content was 48.7%.

TcM2 had a total of 379,716 basecalled reads. All reads were above Q7. The total yield was 1.1 Gbp, with an average sequence length of 2,849 bp, and average quality score of 9.97. The longest read was 51,851 bp in length, and the N50 was 2.78 Kbp. GC content was 43.4%.

TcM3 had a total of 16,536 basecalled reads. 6 reads were below Q7 and removed from further analysis. The total yield was 53.2 Mbp, with an average sequence length and quality score of 3,219 bp and 12.27, respectively. The longest contig was 42,848 bp in length, and the N50 was 2.86 Kbp. GC content was 43.8%.

TcM4 had a total of 3,273,925 basecalled reads. 729,108 reads were below Q7 and removed from further analysis. The total yield was 3.0 Gbp, with an average sequence length, average quality score of 931 bp, and 9.9, respectively. The longest read was 11,318 bp in length, and the N50 was 1.22 Kbp. GC content was 46.1%.

4.3.2 WIMP read classification

Following WIMP analysis on TcM1-3 combined, 255,176 reads were classified as bacteria (247,329, 97%), virus (3,951, 1.5%), eukaryote (3,763, 1.5%) or archaea (133, <1%). 59,259 reads were unclassified.

Individual analysis of TcM1 had 12,578 reads classified as bacteria (9,056, 72%), virus (2,515, 20%), eukaryote (881, 7%) or archaea (126, 1%). 4,577 reads were unclassified. TcM2 had 231,797 reads classified as bacteria (227,805, 98%), virus (1,388, <1%), eukaryote (2,598, 1%) or archaea (6, <1%). 52,022 reads were unclassified. TcM3 had 10,801 reads classified as bacteria (10,468, 97%), virus 48, <1%), eukaryote (284, 3%) or archaea (1, <1%). 2,660 reads were unclassified (Table 4.1).

TcM4 had 517,373 reads classified as bacteria (452,952, 88%), virus (413, <1%), eukaryote (59,284, 12%) or archaea (187, <1%). 2,018,331 reads were unclassified (Table 4.1). TcM4 was not included in microbiome analysis because the PCR library preparation artificially favours specific sequences the primers are targeted towards during amplification and is not a true representation of DNA coverage present in the samples.

Following BLASTn of *H. sapiens* classified reads against NCBI, 612/1567 (39.06%) sequences found a match. Most sequences belonged to various nematode species, particularly *H. contortus* (296/612), *T. circumcincta* (159/612), *Ostertagia* spp. (28/612), and *Chabaudstrongylus ninhae* (25/612), among others (Figure 4.1). Additionally, some bacteria (12/612), bacteriophages (4/612) and mammalian species (55/612; *H. sapiens*, *Ovis aries*, *Odocoileus virginianus* (white-tailed deer)) were also present.

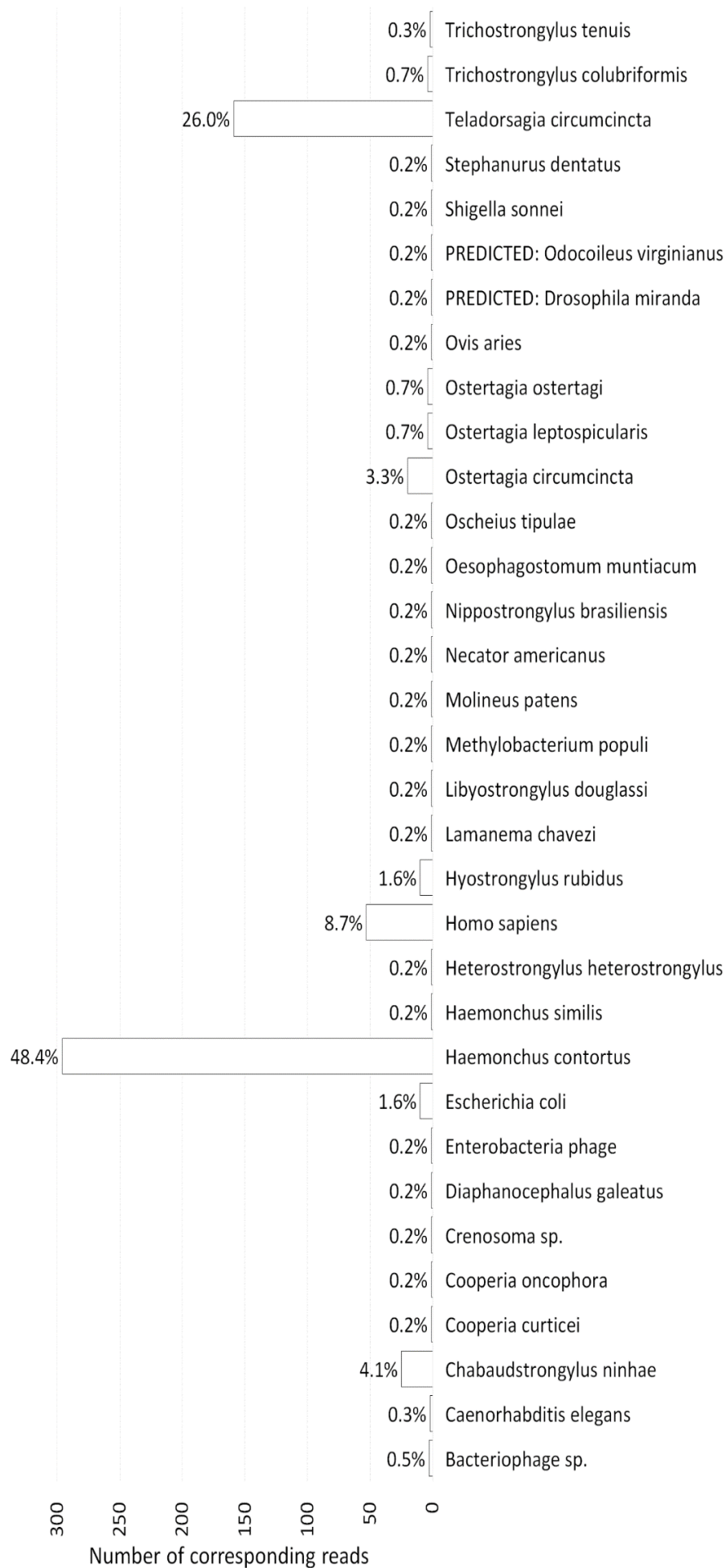


Figure 4.1: Bar graph of the number of reads for each species present in "What's In My Pot?" *Homo sapiens* classified reads which were BLASTn against the NCBI nucleotide database. Percent of overall read total indicated.

4.3.3 Microbiome analysis of TcM1-3 classified reads

The WIMP workflow generated a taxonomic tree of 314,435 reads classified to 24 different phyla. 11/24 phyla appear in all TcM1-3. These 11 phyla include Proteobacteria, Chordata, Actinobacteria, Ascomycota, Firmicutes, Bacteroidetes, Basidiomycota, Euryarchaeota, Mucoromycota, Microsporidia, and Chytridiomycota (Table 4.2) and can be further broken down into 39 classes. 19/39 classes appear in all TcM1-3. These 19 classes include Gammaproteobacteria, Alphaproteobacteria, Betaproteobacteria, Mammalia, Actinobacteria, Saccharomycetes, Leotiomyces, Sordariomycetes, Orbiliomycetes, Eurotiomycetes, Dothideomycetes, Schizosaccharomycetes, Bacilli, Clostridia, Agaricomycetes, Tremellomycetes, Microbotryomycetes, Ustilaginomycetes, and Chytridiomycetes (Table 4.2).

4.3.4 Genome assembly

The final genome assembly of ~114.3 Mbp consisted of 9,219 contigs with 35.45X coverage, a minimum and maximum sequence length of 1,074 and 176,649 bp, respectively (Figure 4.2), and a GC content of 43.7%. We predicted a total of 28,588 protein-coding genes.

Table 4.2: Presence of phyla and classes present in TcM1-3 classified sequence reads

Phylum	Class	TcM 1	TcM 2	TcM 3
Actinobacteria		+	+	+
	Actinobacteria	+	+	+
Ascomycota		+	+	+
	Dothideomycetes	+	+	+
	Eurotiomycetes	+	+	+
	Leotiomyces	+	+	+
	Orbiliomycetes	+	+	+
	Pezizomycetes	+	+	-
	Pneumocystidomycetes	+	+	-
	Saccharomycetes	+	+	+
	Schizosaccharomycetes	+	+	+
	Sordariomycetes	+	+	+
	Xylonomycetes	+	+	-
Aquificae		-	+	-
	Aquificae	-	+	-
Bacteroidetes		+	+	+
	Bacteroidia	+	-	+
	Cytophagia	+	+	-
	Flavobacteriia	+	+	-
	Rhodothermia	+	+	-
	Sphingobacteriia	+	+	-

Phylum	Class	TcM 1	TcM 2	TcM 3
Basidiomycota		+	+	+
	Agaricomycetes	+	+	+
	Exobasidiomycetes	-	+	+
	Malasseziomycetes	-	+	-
	Microbotryomycetes	+	+	+
	Mixiomycetes	+	+	-
	Pucciniomycetes	-	+	+
	Tremellomycetes	+	+	+
	Ustilaginomycetes	+	+	+
	Wallemiomycetes	-	+	-
Caldiserica		-	+	-
	Caldiserica	-	+	-
Chlamydiae		-	+	-
	Chlamydiia	-	+	-
Chlorobi		+	-	-
	Chlorobia	+	-	-
Chordata		+	+	+
	Mammalia	+	+	+
Chytridiomycota		+	+	+
	Chytridiomycetes	+	+	+
Crenarchaeota		+	-	-
	Thermoprotei	+	-	-
Cyanobacteria		+	+	-
	Cyanophyceae	+	+	-
Deinococcus- Thermus		-	+	-
	Deinococci	-	+	-
Elusimicrobia		+	+	-
	Endomicrobia	+	+	-
Euryarchaeota		+	+	+
	Methanobacteria	+	+	-
	Methanococci	-	+	-
	Methanomicrobia	-	+	+
	Thermococci	-	+	-
Firmicutes		+	+	+
	Bacilli	+	+	+
	Clostridia	+	+	+
	Erysipelotrichia	-	+	-
Fusobacteria		-	+	-
	Fusobacteriia	-	+	-
Microsporidia		+	+	+
Mucoromycota		+	+	+
Nitrospirae		-	+	-
	Nitrospira	-	+	-

Phylum	Class	TcM 1	TcM 2	TcM 3
Proteobacteria		+	+	+
	Alphaproteobacteria	+	+	+
	Betaproteobacteria	+	+	+
	Deltaproteobacteria	-	+	-
	Epsilonproteobacteria	+	+	-
	Gammaproteobacteria	+	+	+
Spirochaetes		-	+	-
	Spirochaetia	-	+	-
Synergistetes		-	+	-
	Synergistia	-	+	-
Tenericutes		+	+	-
	Mollicutes	+	+	-

Bold: Present in all TcM1-3.

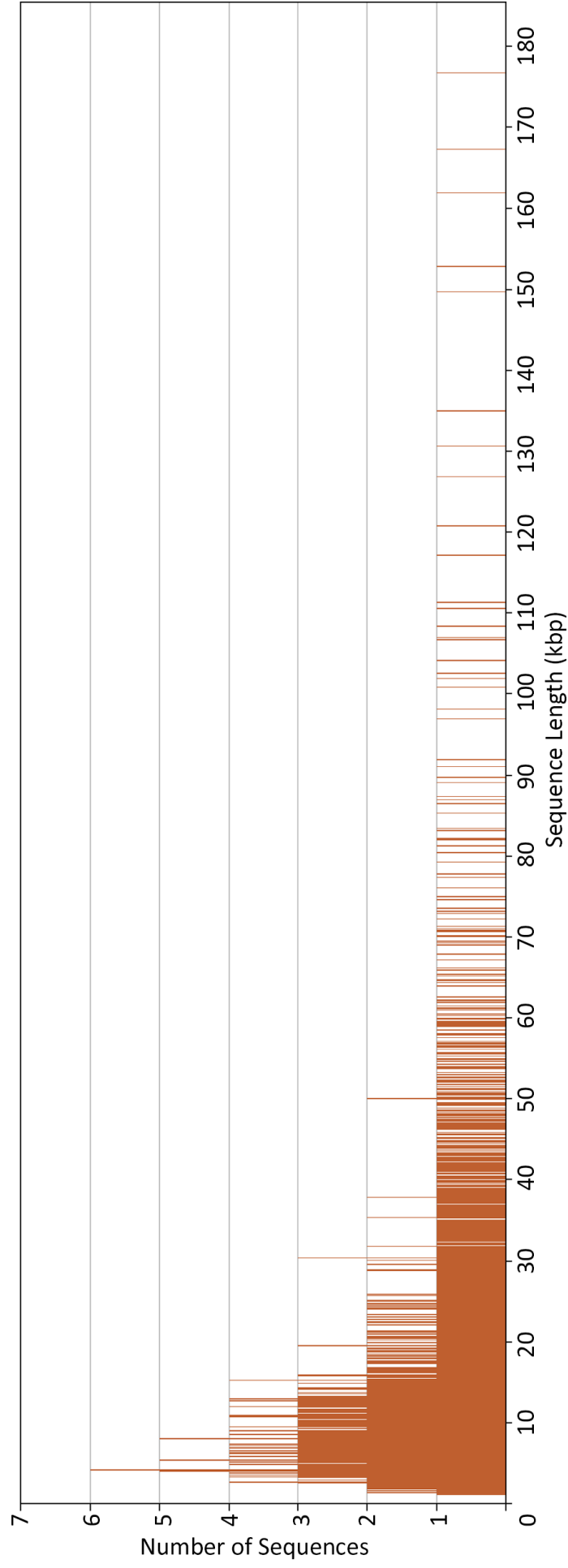


Figure 4.2: Bar graph indicating the length (kbp) of 9,129 sequences in the *Teladorsagia circumcincta* genome assembly in this study, and the number of occurrences for that size.

4.3.5 Cathepsin F gene

A complete *T. circumcincta* cathepsin F gene was found on 2 contigs; contig00000245 (tig245) and contig00000282 (tig282). Tig245 is 59,445 bp in length, made up of 216 reads, and has a GC content of 45.9%. A complete *T. circumcincta* cathepsin F gene is present on this contig with a coding length of 1,095 bp over 10 exons, and a total length of 11,564 bp. An additional 2 incomplete genes are also present; the first excluding exons 3 and 10, and the second made up of only exons 1 – 4 (Figure 4.3).

Tig282 is 34,165 bp in length, made up of 73 reads, and has a GC content of 45.9%. A complete *T. circumcincta* cathepsin F gene is present on this contig with a coding length of 1,101 bp over 10 exons, and a total length of 9,663 bp. An additional incomplete gene is also present made up of only exons 1 – 3 (Figure 4.4).

Multiple sequence alignment of Tci-CF-1 and the tig245 and tig282 complete cathepsin F predicted genes produced high consensus (Figure 4.5). Tig245 substituted residues E₃₀ > D₃₀, K₅₆ > N₅₆, R₇₆ > K₇₆, M₃₄₅ > I₃₄₅, and R₃₅₃ > G₃₅₃. Tig282 inserted residues L₇₁ and V₇₂, and substituted residues E₃₀ > D₃₀, K₅₆ > N₅₆, R₇₆ > K₇₈, R₁₃₈ > Q₁₄₀, P₂₃₅ > S₂₃₇, L₃₀₆ > P₃₀₈, M₃₄₅ > I₃₄₇, and R₃₅₃ > G₃₅₅. The translations of the incomplete genes contain more substitutions (Figure 4.5).

Augustus predicted 21 genes on tig245, and 9 genes on tig282. BLASTn analysis showed these genes matched with Strongyloidea clade V species *H. contortus*, *O. ostertagia*, and *T. circumcincta* with a minimum 66.3% identity, and were mostly complimentary, with predicted genes surrounding the complete cathepsin F annotation matching across the two contigs (Table 4.3). The Augustus predicted genes partially match the annotated cathepsin F gene (Figure 4.3, 4.4).

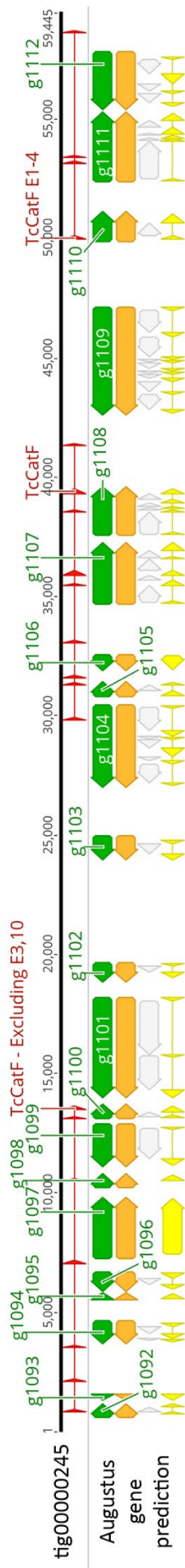


Figure 4.3: Contig00000245 with Augustus predicted genes (g1092 – 1112) and annotated to include *Teladorsagia circumcincta* cathepsin F gene fragments and gene. Arrow shapes indicate the read direction; white: introns; green: gene; yellow: CDS; orange: transcript; red: cathepsin F gene (TcCatF) and gene fragment (full gene excluding exons 3, 10; and exons 1-4) annotations.

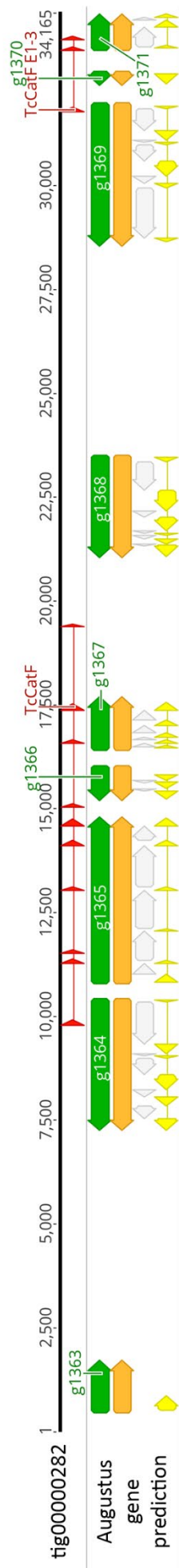


Figure 4.4: Contig00000282 with Augustus predicted genes (g1363 – 1371) and annotated to include *Teladorsagia circumcincta* cathepsin F gene fragments and gene. Arrow shapes indicate the read direction; white: introns; green: gene; yellow: CDS; orange: transcript; red: cathepsin F gene (TcCatF) and gene fragment (exons 1-3) annotations.

	1	10	20	30	40	50	60
Tci-CF-1	MSLLFLLLI	IP	HLFAATVKQ	QYSGGVKPLTE	LRTDLIDKKT	KGSIEFARLG	QHISPKDFGA
Tig245_1DN
Tig282_1DN
Tig245_2VYSRYNELN
Tig245_3	·FVW·F·MS	PS·V·S·K	H·EA·D	PH·D	-----	-----YT·S	RY·N·V·
Tig282_2DN
	70	80	90	100	110	120	
Tci-CF-1	WNHFTSFIER	--HDKVYRNE	SEALKRFGIF	KRNLEIIRSA	QENDKGTAIV	GINQFADLSP	
Tig245_1K	
Tig282_1	LV·K	
Tig245_2	·Q·N	-----	-----	-----TM	EQ·V	TR	
Tig245_3	·Q	---G	RV	····K	····QL·V	····TR	
Tig282_2	---K	
	130	140	150	160	170	180	
Tci-CF-1	EEFKKTHLPH	TWKQPDHPNR	IVDLAAEGVD	PKEPLPESFD	WREHGAVTKV	KTEGHCAACW	
Tig245_1	
Tig282_1Q	
Tig245_2Y	····SH·S	····N·V	····K	····K	····E	····GS
Tig245_3	
Tig282_2	
	190	200	210	220	230	240	
Tci-CF-1	AFSVTGNIEG	QWFLAKKKLV	SLSAQQLLDC	DVVDEGCNGG	FPLDAYKEIV	RMGGLEPEDK	
Tig245_1	
Tig282_1S	
Tig245_2	··T	····R·E	····E·E	····E·W	····L·I	····M·I	····S·D
Tig245_3	
Tig282_2	
	250	260	270	280	290	300	
Tci-CF-1	YPYEAKAEQC	RLVPSDIAVY	INGSVELPHD	EKMRAWLVK	KGPISIGITV	DDIQFYKGGV	
Tig245_1	
Tig282_1	
Tig245_2	····E·K	H·S	··S	····N	····NA	····NMI·S·I
Tig245_3	
Tig282_2	
	310	320	330	340	350	360	
Tci-CF-1	SRPTTCRLSS	MIHGALLVG	YGVGEKNIPYWI	IKNSWGPNWG	EDGYRMRVRG	ENACRINRFP	
Tig245_1I	····G	
Tig282_1PI	····G	
Tig245_2	A·RF·DPDE	LN·V	··I·GKK	····SD	··G	
Tig245_3	
Tig282_2	
	366	364					
Tci-CF-1	TS	AVVL					
Tig245_1					
Tig282_1					
Tig245_2					
Tig245_3					
Tig282_2					

Figure 4.5: Multiple sequence alignment of *Teladorsagia circumcincta* secreted cathepsin F (Tci-CF-1, GenBank accession: ABA01328), and the cathepsin F genes identified in this study.

Tig245_1: complete gene on tig245; Tig282_1: complete gene on tig282; Tig245_2: partial cathepsin F gene on tig245 excluding exons 3 and 10; Tig245_3: partial cathepsin F gene (exons 1 – 4) on tig245; Tig282_2: partial cathepsin F gene (exons 1 – 3) on tig282. Conserved residues indicated by a dot, gaps indicated by a dash.

Table 4.3: Results of BLASTn of Augustus predicted genes on contigs 245 and 282.

Contig	Gene ID	% ID	E-value	Bitscore	Description	Scientific Name	Accession
245	g1092	77.391	6.61E-14	91.5	<i>Teladorsagia circumcincta</i> secreted cathepsin F (cf1) mRNA, complete cds	<i>Teladorsagia circumcincta</i>	DQ133568.1
245	g1093	-	-	-	-	-	-
245	g1094	92.105	0.008	56.3	<i>Haemonchus contortus</i> strain NZ_Hco_NP chromosome 5	<i>Haemonchus contortus</i>	CP035803.1
245	g1095	80.537	2.30E-28	139	<i>Teladorsagia circumcincta</i> secreted cathepsin F (cf1) mRNA, complete cds	<i>Teladorsagia circumcincta</i>	DQ133568.1
245	g1096	81.481	5.58E-23	123	<i>Teladorsagia circumcincta</i> astacin metalloprotease (DPY-31) gene, complete cds	<i>Teladorsagia circumcincta</i>	KM272923.1
245	g1097	72.465	0	1242	<i>Haemonchus contortus</i> , ISE/inbred ISE, WGS project CAVP01000000 data, chromosome: 1	<i>Haemonchus contortus</i>	LS997562.1
245	g1098	78.351	1.22E-10	81.5	<i>Haemonchus contortus</i> strain NZ_Hco_NP chromosome 5	<i>Haemonchus contortus</i>	CP035803.1
245	g1099	66.541	1.13E-29	146	<i>Haemonchus contortus</i> , ISE/inbred ISE, WGS project CAVP01000000 data, chromosome: X	<i>Haemonchus contortus</i>	LS997567.1
245	g1100	88.636	6.20E-21	114	<i>Haemonchus contortus</i> strain NZ_Hco_NP chromosome 5	<i>Haemonchus contortus</i>	CP035803.1
245	g1101	75	1.66E-06	69.8	<i>Ostertagia circumcincta</i> galectin GAL-1 gene, complete cds	<i>Teladorsagia circumcincta</i>	AF105337.1
245	g1102	70.404	1.49E-17	104	<i>Teladorsagia circumcincta</i> clone MTG1a microsatellite sequence	<i>Teladorsagia circumcincta</i>	DQ355411.1
245	g1103	69.406	1.01E-14	95.1	<i>Haemonchus contortus</i> strain NZ_Hco_NP chromosome X	<i>Haemonchus contortus</i>	CP035800.1
245	g1104	66.951	1.18E-51	220	<i>Haemonchus contortus</i> , ISE/inbred ISE, WGS project CAVP01000000 data, chromosome: 5	<i>Haemonchus contortus</i>	LS997566.1
245	g1105	97.802	2.09E-33	156	<i>Teladorsagia circumcincta</i> secreted cathepsin F (cf1) mRNA, complete cds	<i>Teladorsagia circumcincta</i>	DQ133568.1

Contig	Gene ID	% ID	E-value	Bitscore	Description	Scientific Name	Accession
245	g1106	75.758	1.06E-05	64.4	<i>Haemonchus contortus</i> , ISE/inbred ISE, WGS project CAVP01000000 data, chromosome: 3	<i>Haemonchus contortus</i>	LS997564.1
245	g1107	87.77	0	866	<i>Teladorsagia circumcincta</i> clone MTG1a microsatellite sequence	<i>Teladorsagia circumcincta</i>	DQ355411.1
245	g1108	99.451	2.29E-83	324	<i>Teladorsagia circumcincta</i> secreted cathepsin F (cf1) mRNA, complete cds	<i>Teladorsagia circumcincta</i>	DQ133568.1
245	g1109	70.871	0	1366	<i>Haemonchus contortus</i> , ISE/inbred ISE, WGS project CAVP01000000 data, chromosome: 5	<i>Haemonchus contortus</i>	LS997566.1
245	g1110	73.188	7.79E-11	83.3	<i>Haemonchus contortus</i> strain NZ_Hco_NP chromosome 1	<i>Haemonchus contortus</i>	CP035805.1
245	g1111	66.324	7.22E-41	183	<i>Haemonchus contortus</i> , ISE/inbred ISE, WGS project CAVP01000000 data, chromosome: X	<i>Haemonchus contortus</i>	LS997567.1
245	g1112	84.239	0	1003	<i>Haemonchus contortus</i> , ISE/inbred ISE, WGS project CAVP01000000 data, chromosome: 5	<i>Haemonchus contortus</i>	LS997566.1
282	g1363	85.038	0	1014	<i>Teladorsagia circumcincta</i> clone MTG12 microsatellite sequence	<i>Teladorsagia circumcincta</i>	DQ355421.1
282	g1364	66.951	1.08E-51	220	<i>Haemonchus contortus</i> , ISE/inbred ISE, WGS project CAVP01000000 data, chromosome: 5	<i>Haemonchus contortus</i>	LS997566.1
282	g1365	98.765	3.93E-71	284	<i>Teladorsagia circumcincta</i> secreted cathepsin F (cf1) mRNA, complete cds	<i>Teladorsagia circumcincta</i>	DQ133568.1
282	g1366	82.857	2.67E-08	73.4	<i>Haemonchus contortus</i> strain NZ_Hco_NP chromosome 5	<i>Haemonchus contortus</i>	CP035803.1
282	g1367	98.901	1.74E-82	320	<i>Teladorsagia circumcincta</i> secreted cathepsin F (cf1) mRNA, complete cds	<i>Teladorsagia circumcincta</i>	DQ133568.1
282	g1368	69.919	2.76E-140	513	<i>Haemonchus contortus</i> , ISE/inbred ISE, WGS project CAVP01000000 data, chromosome: 2	<i>Haemonchus contortus</i>	LS997563.1
282	g1369	66.951	1.18E-51	220	<i>Haemonchus contortus</i> , ISE/inbred ISE, WGS project CAVP01000000 data, chromosome: 5	<i>Haemonchus contortus</i>	LS997566.1

Contig	Gene ID	% ID	E-value	Bitscore	Description	Scientific Name	Accession
282	g1370	94.643	4.30E-13	88.7	<i>Ostertagia ostertagi</i> cathepsin B-like cysteine protease gene, partial cds	<i>Ostertagia ostertagi</i>	M88505.1
282	g1371	97.802	2.94E-33	156	<i>Teladorsagia circumcincta</i> secreted cathepsin F (cf1) mRNA, complete cds	<i>Teladorsagia circumcincta</i>	DQ133568.1

Matching predicted genes across contigs highlighted.

4.4 Discussion

4.4.1 Genome assembly

This study is the first to use Oxford Nanopore Technology to sequence *T. circumcincta* DNA, though it is not the first to assemble a *T. circumcincta* genome. The estimated genome size from this assembly is ~114.3 Mbp, comprises 9,219 contigs, and has an estimated 28,588 protein-coding genes. This quantity of protein coding genes is similar to the draft genome for *T. circumcincta* developed by Choi *et al.* (2017) using partially in-bred drug-susceptible strains to compare genome-wide single nucleotide and copy number variants of drug-resistant *T. circumcincta* strains (BioProject PRJNA72569). The estimated *T. circumcincta* genome size from their assembly was ~701 Mbp, comprised 81,730 contigs, has an estimated 25,532 protein-coding genes, and utilised short-read sequencing technology (Choi *et al.*, 2017).

Additionally, The Wellcome Trust Sanger Institute undertook *T. circumcincta* genome sequencing (BioProject PRJEB7676) using predominantly PacBio SMRT sequencing, and some Illumina HiSeq sequencing on pooled L3 *T. circumcincta*. 240 Gbp of total data is available through the Sequence Read Archive, however, there is no indication that this data has been used in an assembly as yet.

Genome assembly tools and programs often require a genome size estimate for construction and a correct assembly requires minimal gaps, no unnecessary repeat regions, and most importantly, the sequence in the correct order. Determining the size of an organism's genome is difficult, but, flow cytometry has been considered a fast, sensitive technique for determining genome sizes in many organisms (Nath *et al.*, 2014; J. Wang *et al.*, 2015). Leroy *et al.* (2003) used flow cytometry to estimate genome sizes of a range of nematode species, using the known *Caenorhabditis elegans* genome as calibration. The estimated genome size for *T. circumcincta* was calculated as 58.6 Mbp, and this size was used for the assembly parameters in the current study. *C. elegans* is the model organism for nematode species, has been extensively researched and belongs to Order Rhabditida alongside *T. circumcincta*. The genome of *C. elegans* was the first nematode species to be sequenced completely and is 100 Mbp (Wood, 2002a), with 19,735 protein-coding genes (Hillier *et al.*, 2005).

The *C. elegans* genome was also used to predict genes in our assembly. The predicted genes on tig245 and tig282 did not match well to our manual annotation of cathepsin F. This was also seen in the Choi *et al.* (2017) assembly and was previously discussed in Chapter 2 (Sloan *et al.*, 2020). If a species of closer genealogy were to be used as the reference, the predicted genes may be more accurate. *Necator americanus*, for example, is a closer relative to *T. circumcincta* than *C. elegans*. Unfortunately, *N. americanus* is not available for reference in the version of Augustus gene prediction software that was used in this study. Despite not matching correctly to Tci-CF-1, the

predicted genes detected on tig245 and tig282 all had BLAST hits to nematode species genes. This is a positive indication that Augustus is correctly predicting genes specific for nematode species. Of particular interest is that the predicted genes match those of *Haemonchus contortus*, *Ostertagia ostertagi*, *O. circumcincta* (the previous name for *T. circumcincta*) and *T. circumcincta*. All these nematode species are closely related to each other and occur in the same subfamilies called Ostertagiinae or Haemonchidae.

Two complete cathepsin F genes of 9,663 bp and 11,564 bp in length and each with 10 exons were identified in this study. The predicted protein sequences were very similar to the Tci-CF-1 sequence. Sloan *et al.* (2020) used bioinformatic analyses to explore the cathepsin F gene of *T. circumcincta* and estimated the Tci-CF-1 gene to be a minimum of 9,583 bp in length (and predicted it would be longer as it contained two gap regions) with 10 exons. The analysis in this study provides further confidence that the gene construction is correct as gaps were filled and the encoded protein is highly similar. Our construction also indicates that there are two different regions encoding for cathepsin F, with slight variation between them. Due to the high consensus of the surrounding predicted genes across the two contigs it is likely that these are different alleles, as *T. circumcincta* is a diploid organism, rather than distinct genes. The additional gene fragments on tig245 and tig282 are highly similar to Tci-CF-1 and the predicted complete sequences. Perhaps this variation allows for increased host immune evasion by slightly altering the protein structure and requiring different conformation for bonding. Interestingly, all substitutions from Tci-CF-1 present in this study on both tig245 and tig282 were identified previously by Sloan *et al.* (2020) as potential polymorphic variants.

4.4.2 Microbiome analysis

The construction of the *T. circumcincta* genome is a difficult endeavour, especially for ensuring the data used for construction is of the correct species. It is particularly difficult to obtain a “clean” DNA extraction from gut parasites because of their habitat. Because *T. circumcincta* cannot be grown from egg to adult *in vitro*, it is especially difficult to ensure contaminants are not present. Consequently, removal of contaminating reads post-sequencing is necessary. Our read classification has shown that even with thorough washing of specimens, other species remain. *T. circumcincta* is topologically a tube and cleansing the inside of contaminants is difficult. It could be that the bacteria and fungi present in or attached to *T. circumcincta* are crucial for survival of the nematode within the host, and this could explain why *in vitro* culture has so far been unsuccessful.

The 3 kingdoms present in all TcM1-3 were primarily Bacteria, Fungi and Eukaryota. The Eukaryotic reads are likely those of *T. circumcincta* because WIMP is programmed to select *H. sapiens* for reads which match any eukaryotic species that are not fungi. It is crucial to explore further which

species these reads actually belong to. Some reads were correctly identified as *H. sapiens*, however, the rest were predominantly gastrointestinal nematode species. For example; *Chabaudstrongylus ninhae* is found in the gut of Reeve's muntjacs (Setsuda *et al.*, 2019), *Libyostrongylus douglassi* in the gut of ostriches (Sanchez-Ayala *et al.*, 2018), *Lamanema chavezii* in the gut of alpacas and llamas (Cafrune *et al.*, 2001), *Cooperia oncophora* in the gut of cattle and shown to often co-infect with *O. ostertagi* and *H. contortus* (R. W. Li *et al.*, 2011), and *Trichostrongylus colubriformis* in the gut of sheep and goats (Woolastont *et al.*, 2001). The sheep in this study were drenched prior to experimental infection with *T. circumcincta*, and faeces were examined to confirm the absence of nematode species, so it is unlikely that the presence of these other nematode species is accurate. It is more likely that these nematode species have genomic sequences that are similar to *T. circumcincta*. They may have been misassigned because the *T. circumcincta* genome lacks the complete gene sequence or there are errors in the sequence. Oxford Nanopore Technology, and the MinION™, were previously known to have high error rates of ~12%, distributed amongst mismatches, insertions and deletions (Kchouk *et al.*, 2017), but this has dropped to ~5% in recent years (Amarasinghe *et al.*, 2020).

It is surprising that only one read matched with *O. aries*. The expectation is there would be a greater presence of host DNA in a sample. Additionally, the presence of a single white-tailed deer read is both odd but not entirely unexpected as feral deer are common in Victoria, Australia, and pasture may contain deer contaminants. Alternatively, the DNA sequenced belongs to sheep but shares similarities with deer, and incomplete sequences or sequencing error has resulted in incorrect classification. It is possible that the preparatory wash steps prior to DNA extraction were able to remove contaminants from the outside of the worms, but not the inside, and the additional species sequenced in this study were present in the gut of the *T. circumcincta* specimens. Bleach has been shown to be an effective wash method of specimens, and does not interfere with DNA extraction from maggots, teeth and bones (Kemp *et al.*, 2005; Linville *et al.*, 2002). It may be worthwhile incorporating bleach-sterilisation into sample preparation in future.

Bacterial and fungal relationships with *T. circumcincta* are of particular interest because of potential competitive, commensal, mutual or parasitic interactions. The majority of classified reads belonged to Bacteria and those found in all TcM1-3 were phylum Proteobacteria, Actinobacteria, Firmicutes, and Bacteroidetes.

Proteobacteria are Gram-negative bacteria that include a wide variety of pathogenic and non-pathogenic species. Some species are agriculturally important as they are capable of inducing nitrogen fixation in symbiosis with plants (Williams *et al.*, 2007), and denitrification of waste-water (Bonnet *et al.*, 1997; Cydzik-Kwiatkowska *et al.*, 2016). Interestingly, the higher the soil pH, the higher the relative abundance of Alpha-, Beta-, and Gammaproteobacteria (Rousk *et al.*, 2010).

Gamma proteobacteria are the most phylogenetically and physiologically diverse class of Proteobacteria and contain genera such as *Escherichia*, *Shigella*, *Salmonella*, *Yersinia*, and *Pseudomonas* etc. All of which were detected in the classified reads, and the overwhelming majority of which were classified as *Escherichia spp.*, specifically *E. coli*. This is unsurprising as *E. coli* are well known to inhabit the gastrointestinal tracts of animals (Silva *et al.*, 2012). Alphaproteobacteria can grow at very low nutrient levels and comprise species capable of nitrogen fixation for plants (Brenner *et al.*, 2005; Williams *et al.*, 2007). Additionally, the mitochondria of eukaryotes are thought to be descendants of Alphaproteobacteria (Roger *et al.*, 2017). Betaproteobacteria are useful in wastewater treatment systems because of their ability to denitrify and remove excess ammonia (Bonnet *et al.*, 1997; Cydzik-Kwiatkowska *et al.*, 2016).

Actinobacteria are mostly Gram-positive bacteria and contribute to soil systems. They help decompose dead organisms and organic matter for plant nutrient uptake (Ventura *et al.*, 2007). Some species fix nitrogen for plants in exchange for access to the plant's sugars. In agriculture, Actinobacteria are used as insecticides, herbicides, fungicides, and growth-promoting substances for plants and animals (Gupte *et al.*, 2002; Mahajan *et al.*, 2012). The presence of Actinobacteria in the abomasum of sheep is expected from grazing behaviours.

Firmicutes are Gram-positive bacteria which play an important role in beer, wine, and cider spoilage. They are known to make up the largest portion of mouse and human gut microbiomes (Ley *et al.*, 2006), as well as the rumen of ruminants (Callaway *et al.*, 2010; Chaucheyras-Durand *et al.*, 2014; Kamra, 2005; Lee *et al.*, 2012; Lopes *et al.*, 2015; McLoughlin *et al.*, 2020; Perumbakkam *et al.*, 2011; Pitta *et al.*, 2010). As part of the gut flora, they have been shown to be involved in energy resorption. Firmicute classes found in all TcM1-3 are Bacilli and Clostridia. Clostridia are distinguished from Bacilli by respiration. Clostridia are obligate anaerobes while Bacilli perform aerobic respiration. Most *Clostridium* species ferment plant polysaccharides and are found in soil, however, some species are found in human and animal microbiota (Jalanka *et al.*, 2018). Bacilli are found in soil, water, plants, animals and humans. Most species are nonpathogenic, however, some species are capable of causing disease in animals and humans by toxin production (Delbrassinne *et al.*, 2016).

Bacteroidetes are Gram-negative bacteria that are widely present in the environment including in the gut and on the skin of animals (Rajilic-Stojanovic *et al.*, 2014). Many Bacteroidetes species are symbiotic and adapted to the gastrointestinal tract where they perform metabolic conversions that are essential for the host such as protein and sugar degradation (Thomas *et al.*, 2011). No one class of Bacteroidetes was present in all TcM1-3.

These bacterial phyla and classes are likely to be present in pasture, soil, water, or host gut. Their presence in the sheep gut, and therefore presence in and around *T. circumcincta*, is expected and may easily be extracted with *T. circumcincta* DNA. Additionally, the fungal phyla found in all TcM1-3 (Ascomycota, Basidiomycota, Chytridiomycota, and to a lesser extent Mucoromycota and Microsporidia) would also be expected as they are all either decomposers, associated with plants, or associated with parasites, and are likely to have been present on pasture and therefore in the sheep gut alongside *T. circumcincta*.

Ascomycota are the largest phylum of Fungi, with over 64,000 species and are commonly known as sac fungi (Lutzoni *et al.*, 2004; Rivera-Mariani *et al.*, 2011). Common examples include truffles, brewer's and baker's yeasts, and *Penicillium* (Mello *et al.*, 2006; Mortimer *et al.*, 1999). There are several species which are plant pathogens, including powdery mildews, apple scab, and rice blast (Schoch *et al.*, 2009). The specific classes of Ascomycota present in all TcM1-3 are Orbiliomycetes, Saccharomycetes, Schizosaccharomycetes, Sordariomycetes, Leotiomyces, Eurotiomycetes, and Dothideomycetes. Of particular interest are the Orbiliomycetes which are well known for carnivorous and nematophagous fungal species that are common in temperate regions, where *T. circumcincta* is most prominent. These Orbiliaceae have evolved specialised mechanisms to trap nematodes. Fungal mycelia penetrate the nematode and differentiate into trap structures; adhesive networks, columns and knobs, and constricting and non-constricting rings, which ultimately digest the nematode's internal contents (Yang *et al.*, 2007).

Saccharomycetes and Schizosaccharomycetes are decomposer yeasts primarily feeding on decaying wood, leaves and other organic matter, and have associations with insects, plants and humans (Alexopoulos *et al.*, 1996; Kirk *et al.*, 2008). Sordariomycetes are decomposers known to grow in animal faeces, soil, and decaying organic matter. Leotiomyces cause serious plant disease and are considered a sister taxon to Sordariomycetes. Eurotiomycetes are the *Penicillium/Aspergillum* fungi and are common everywhere. Dothideomycetes are plant pathogens that grow on woody debris, decaying leaves, or faeces, and some can form associations with plant roots (Alexopoulos *et al.*, 1996).

Basidiomycota include mushrooms, bracket fungi, rusts, and the human pathogenic yeast *Cryptococcus* (Lutzoni *et al.*, 2004; Rivera-Mariani *et al.*, 2011). The classes present in all TcM1-3 include Agaricomycetes, Tremellomycetes and Ustilaginomycetes. Agaricomycetes are mushrooms which function as decayers of wood and occur in a wide range of environments. Some species are pathogenic or parasitic, and some are symbionts of forest trees (Ingold *et al.*, 1993). The Tremellomycetes are jelly fungi which can be found in plants, humans, and animals (Ingold *et al.*, 1993). Ustilaginomycetes are plant parasite smut fungi (Alexopoulos *et al.*, 1996; Bauer *et al.*, 1997).

Chytridiomycota are zoosporic fungi which can degrade chitin and keratin, and sometimes act as parasites (Lutzoni *et al.*, 2004). They are found in soil and fresh water, parasitise algae, and at least two species have been shown to infect amphibian species (Alexopoulos *et al.*, 1996; Berger *et al.*, 1999).

Mucoromycota are molds and consist of mainly mycorrhizal fungi, root endophytes and plant decomposers (Alexopoulos *et al.*, 1996). No one class of Mucoromycota were present in all TcM1-3.

Microsporidia were once considered protozoan unicellular parasites but are now considered fungi (Hibbett *et al.*, 2007). Microsporidia are restricted to animal hosts often causing chronic disease. They mostly infect insects but have been shown to infect humans, animals, and other parasites (Didier, 2005; Toguebaye *et al.*, 2014). Effects can include reduced fertility, weight, and longevity, and vertical transmission is frequently reported. No one class of Microsporidia were present in all TcM1-3.

The presence of both fungus and bacteria in association with *T. circumcincta* could be a symbiotic relationship where nutrients and amino acids are shared between species for the benefit of both. It may be a reason as to why *in vitro* cultures of *T. circumcincta* are unsuccessful, the presence of these microbes may be essential for ongoing survival. Luque *et al.* (2010) showed the survival of adult *T. circumcincta* specimens *in vitro* was improved by the presence of a mammalian cell line. Perhaps additional microbial factors could enhance this further. In future, sequencing a portion of the environment surrounding the specimens as an additional control, to do a differential analysis with the aim of identifying microbes specifically enriched within the nematode would be beneficial.

4.5 Conclusion

Correct assembly of genomes is difficult, but the present study has produced a draft assembly which has built on existing assemblies (Choi *et al.*, 2017). An assembled genome is valuable for the advancement of knowledge on this parasite. Applications include host treatment, vaccine development or disease control. MinION™ sequencing generated over 3 million reads and these reads were successfully combined to create a draft sequence for cathepsin F analysis. This assembly differed from Choi *et al.* (2017) because of reduced contigs and overall genome size, however, examination of the cathepsin F sequence indicated that correct assembly was achieved for this gene. An in-depth examination of other genes would be beneficial in confirming improved assembly.

In addition, there were several sequences that appeared to be from other species, and this raises questions about the role of the nematode microbiome and its influences on nematode survival

inside and outside the host. These sequences were predominantly bacteria and fungi of species either known to cohabit with *T. circumcincta* or reasonably expected to do so. Very few host DNA sequences were recovered, indicating washing of specimens during DNA extraction is successfully cleansing the outer surface of the specimen. Post-sequencing read analysis and processing has been identified as crucial to remove contaminants.

4.6 Supplementary Information

S4.1 Supplementary File 1: Genome assembly contigs of the *Teladorsagia circumcincta* canu assembly created using Oxford Nanopore Technology reads sequenced in this study and PacBio RS reads retrieved from the Sequence Read Archive (BioProject PRJEB7676). Data available from:

<https://drive.google.com/file/d/1Bcx-7ibmLpk7ZxMETyGwJKUL1iKC5f3i/view>

S4.2 Supplementary File 2: General feature format 3 file of the Augustus gene predictions for the *Teladorsagia circumcincta* assembly created in this study. This file is to be loaded with the assembly in S4.1. Data available from:

https://drive.google.com/file/d/1a_PHeyOwHTzetfuYBkFWsBhxj4HfE9T9/view

Chapter 5

Preliminary cloning and expression of the secreted cathepsin F protein from
Teladorsagia circumcincta larvae

5.1 Introduction

Teladorsagia circumcincta is the most important parasitic nematode of sheep in cool temperate regions worldwide (Bartley *et al.*, 2003) and has a standard one-host lifecycle which has been outlined in Chapter 1. Cathepsin F, a cysteine protease, is the most abundant protein secreted by fourth-stage larvae (L4) and may be crucial for establishment within the host. Clinical disease caused by *T. circumcincta* infection results in reduced production, decreased animal welfare, parasitic gastroenteritis, poor growth performance, and weight loss (Stear *et al.*, 2003).

Cysteine proteases are enzymes influencing processes involving cell death, protein degradation, post-translational modifications of proteins, extracellular matrix remodelling, autophagy, and immune signalling (Dana *et al.*, 2020). Parasitic cysteine proteases are involved in parasite stage transition, invasion of host tissues, nutrient uptake, and immune evasion. In helminth parasites, cysteine proteases are most often secreted externally (T. H. Kang *et al.*, 2004). Mature cysteine proteases are composed of 2 domains; a left and a right domain with the active site cleft inbetween (Vidak *et al.*, 2019). These proteases are characterised by a “catalytic triad” active site; a cysteine, a histidine and an asparagine residue (Deussing *et al.*, 2000), with the asparagine residue orientating the imidazole ring of histidine to form the catalytic triad (Barrett *et al.*, 2001). Cathepsins are synthesised as zymogens (an inactive precursor) with an N-terminal propeptide which inhibits the enzyme action. Cathepsin propeptide inhibitors are α -helical domains which physically prevent access to the substrate-binding active site cleft. Proteolytic cleavage removes the inhibitor and activates the enzyme (Groves *et al.*, 1996). After activation, their proteolytic activity is kept under control by pH, compartmentalisation, and by inhibitors such as cystatins, stefins, kininogens, thyropins, and some serpins (Barrett *et al.*, 2001; Vidak *et al.*, 2019). Cathepsin F is a relatively new cysteine protease in *T. circumcincta* (T. H. Kang *et al.*, 2004; Sloan *et al.*, 2020; B. Wang *et al.*, 1998) and is known to be targeted by IgA antibodies of immune sheep (Nisbet *et al.*, 2013), but the function or role of cathepsin F in *T. circumcincta* is not yet known.

Recombinant expression of the cathepsin F protein would allow for the function or role to be studied. Cathepsins of other species have been recombinantly produced in bacteria: *Escherichia coli* (Novinec *et al.*, 2012; S. M. Smith *et al.*, 1989); and yeasts: *Saccharomyces cerevisiae* (Brömme *et al.*, 1993; Law *et al.*, 2003), and *Pichia pastoris* (Dağlioğlu, 2017; Linnevers *et al.*, 1997; Puzer *et al.*, 2004). The different systems have advantages and disadvantages depending on the purpose of the resulting protein product. For instance, for a long time it was believed *E. coli* was not capable of glycosylation, and the yeast expression systems were necessary to ensure glycosylation of some proteins. However, this problem has now been resolved in *E. coli* (Chen, 2012).

By targeting the cathepsin F protein and understanding its functional role, methods of nematode control could be targeted to the pivotal L4 life stage, to prevent development into adults and

reduce damage. This study aimed to produce active recombinant cathepsin F protein for further work in functional assays. However, due to the advent of Covid-19 this study was abandoned before completion. This chapter acts as a launching platform for the next researcher.

5.2 Materials & Methods

5.2.1 Cloning and transformation of TcCatF cDNA in pPICZ α B

A previous cloning and transformation of TcCatF into *P. pastoris* was performed by a colleague and will be referred to as prev-TcCatF. Cloning and transformation was repeated due to missing/inadequate documentation, and lack of positive results, and the new cloning and transformation will be referred to as new-TcCatF.

Constructs encoding TcCatF were codon-optimised, commercially synthesised by Bioneer Pacific (Korea), and cloned into the secretory *P. pastoris* pPICZ α B vector (Invitrogen). Recombinant TcCatF expression was performed following the EasySelect™ Pichia Expression Kit user manual (Invitrogen, Cat #K1740-01, Carlsbad, USA). Briefly, the *T. circumcincta* cathepsin F pro-mature coding region plasmid DNA (TcCatF) was inserted into *E. coli* DH5 α cells using the heat-shock method as per manufacturer's instructions. Empty pPICZ α B was inserted into *E. coli* Top10 cells using the heat-shock method as per manufacturer's instructions. DH5 α and Top10 cells were transferred to low-salt LB agar plates (15 g/L agar, 0.5 g/L NaCl, 10 g/L tryptone, 5 g/L yeast extract) with Zeocin™ (25 μ g/ml) and grown at 37°C overnight. Several single colonies were selected and streaked again on new low-salt LB agar plates with Zeocin™ (25 μ g/ml) and grown at 37°C overnight. Single colonies were selected and grown in low-salt LB media (0.5 g/L NaCl, 10 g/L tryptone, 5 g/L yeast extract) with Zeocin™ (25 μ g/ml) at 37°C overnight. Plasmids were extracted from *E. coli* DH5 α containing TcCatF and Top10 containing pPICZ α B using AccuPrep® Plasmid Mini Extraction Kit (Bioneer, Cat #K-3030) as per manufacturer's instructions. Restriction digestion was performed to linearise the plasmids using SacI-HF enzyme. Plasmids were run on 0.8% agarose gel electrophoresis to confirm linearization.

P. pastoris X33 and GS115 cells were streaked onto Yeast Peptone Dextrose plates (YPD; 20 g/L agar, 20 g/L dextrose, 20 g/L peptone, 10 g/L yeast extract) and grown at 29°C for 2 days. Separate 10 mL aliquots of YPD media (20 g/L dextrose, 20 g/L peptone, 10 g/L yeast extract) were inoculated with single colonies of GS115 and X33 cells from the previous plates and incubated at 28-30°C, 200 rpm, overnight. Once cells reached an optical density at 600 nm of 0.6-1.0 they were pelleted by centrifugation and prepared for competency under EasyComp™ Transformation.

Linearised TcCatF plasmid DNA was transformed into competent X33 and GS115 cells and grown on YPDS (20 g/L agar, 20 g/L dextrose, 20 g/L peptone, 10 g/L yeast extract, 1 M sorbitol) + Zeocin™ (100 μ g/ml) plates for 3-10 days at 37°C. Single colonies were selected, resuspended in 100 μ L YPD

media, replated on YPD + Zeocin™ (100 µg/ml and 250 µg/ml) plates and grown for 3-10 days at 29°C. Growth was observed on YPD + Zeocin™ (100 µg/ml) plates and colonies were transferred to a YPD + Zeocin™ (100 µg/ml) selection plate. Selection plate was grown for 2-3 days at 29°C.

Successful single colonies were picked and resuspended in 20 µL H₂O. 5 µL aliquots of each colony were boiled at 95°C for 10 minutes. 19 colonies from each X33 and GS115 transformations were amplified by PCR with oligonucleotides:

- AOX1-Forward: 5'TATCGTCGACGGATCAATGGCAACAGTTAAACAACAGTAC
- AOX1-Reverse: 5'TACCTGCAGGGAATTTATAGAACAACCGCTGACGTGGG

Amplification as performed under the following conditions; 5 µL GoTaq Green Master Mix (Promega, M7122), 0.5 µL AOX1-Forward primer (10 µM), 0.5 µL AOX1-Reverse primer (10 µM), 1 µL DNA template, and 3 µL nuclease-free H₂O. The negative control contained nuclease-free H₂O, and the positive control contained linearized empty pPICZα B plasmid DNA. Amplification at 1,582 bp indicates successful transformation of *P. pastoris* cells with TcCatF. Progress with new-TcCatF stopped here.

5.2.2 Preparation of recombinant proteins from yeast

Expression of prev-TcCatF used a successful X33 transformant from a previous transformation (Figure 5.1). Expression of new-TcCatF was not conducted.

Standard TcCatF expression protocol was as follows: A 10 mL starter culture containing prev-TcCatF (X33 *P. pastoris* transformation) in YPD media was grown at 28°C, 180 rpm for 48 hours. In a 2 L baffled flask, expression culture media (320 mL YEP (10 g/L yeast extract, 10 g/L bacto peptone, 5 g/L NaCl), 40 mL KH₂PO₄ (7.35%; pH 6.0), 40 mL Yeast Nitrogen Base (15% w/v), 0.8 mL Biotin (0.02% w/v), 10 mL dextrose (20% w/v), and 1 mL methanol) was inoculated with 5 mL prev-TcCatF starter culture and incubated at 22°C, 160 rpm for 96 hours for expression of recombinant protein. Every 24 hours 2 mL methanol was added to the culture to induce protein expression. Protein expression culture was centrifuged at 6000 x g for 20 mins. The supernatant was stored at 4°C, and pellet discarded.

Yeast culture supernatant containing prev-TcCatF was dialysed against four changes of buffer A (25 mM NaH₂PO₄, 10 mM imidazole, 250 mM NaCl, pH 7.6). Recombinant protein was purified by incubation with Ni-nitrilotriacetic acid (NTA) (Qiagen) in batch format. After washing the resin with 10 bed volumes of buffer A, TcCatF was eluted from Ni-NTA with buffer A containing 250 mM imidazole. The purified protein was stored at -20°C. Concentration of purified protein was measured with Pierce BCA Protein Assay Kit as per manufacturer's instructions (Thermo Fisher Scientific, Catalog #: 23225).

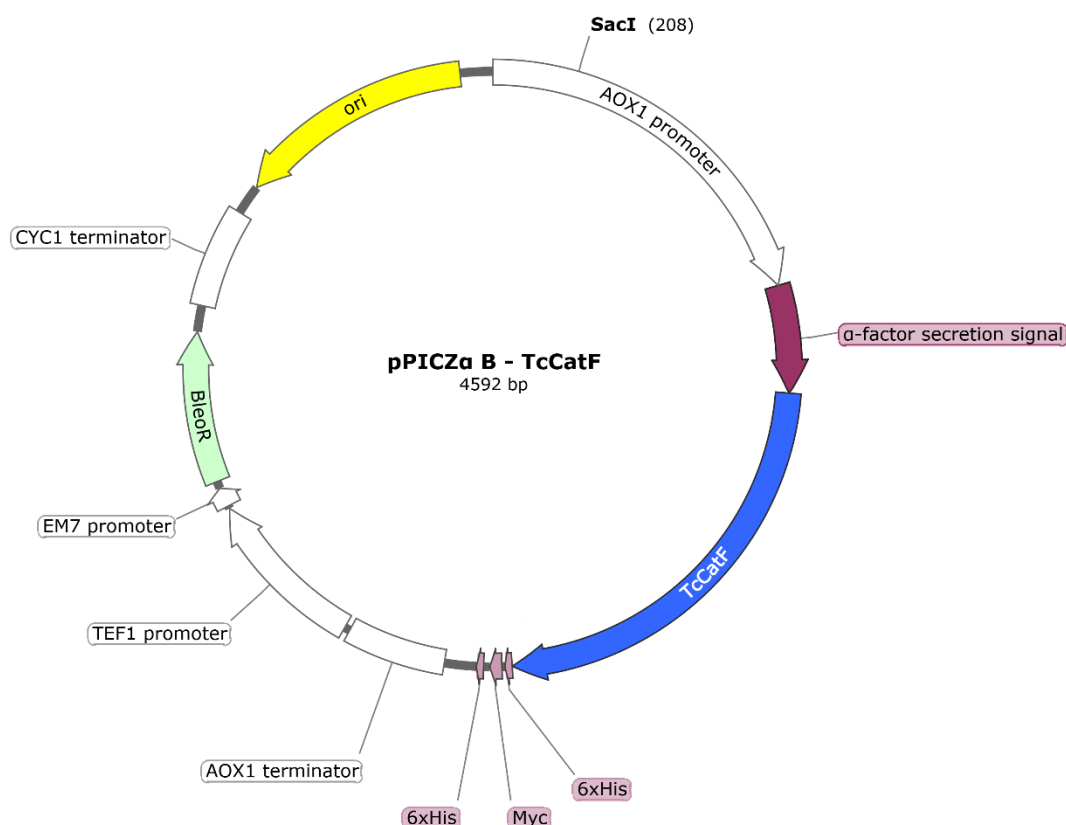


Figure 5.1: Map of the expression vector pPICZα B - TcCatF. AOX1 promoter: Alcohol oxidase 1 promoter region allows methanol-inducible expression in *P. pastoris*, 208 bp; TcCatF: sequence encoding *Teladorsagia circumcincta* secreted cathepsin F; 6xHis: 6x Histamine affinity tag; Myc: human c-Myc proto-oncogene epitope tag; AOX1 terminator: transcription termination region; TEF1 promoter: promoter for EF-1α; EM7 promoter: synthetic bacterial promoter; BleoR: antibiotic-binding protein confers resistance to Zeocin™; CYC1 terminator: transcription termination region processing of resistance gene mRNA; ori: origin of replication functional in *E. coli*; and SacI: SacI/SacI-HF restriction enzyme site.

Eleven attempts of protein expression were documented using prev-TcCatF, with variations. Variations to the standard protocol include starter culture volume, dextrose concentration, incubation time, temperature and rpm, Zeocin™ inclusion and concentration, expression culture volume, and flask type, and are described in Table 5.1.

Table 5.1: Starter and expression culture setup methods for production of prev-TcCatF

Phase	Variable	Standard	1	2	3	4	5	6	7
Starter Culture	YEP Media Volume (mL)	9	9	9	9	9	9	90	90
	Dextrose Volume (mL)	1	1	1	1	1	1	10	10
	prev-TcCatF Inoculation (+/-)	+	+	+	+	+	+	+	+
	Incubation Temp (°C)	28	28	28	28	28	28	28	28
	Incubation RPM	180	180	200	180	160	160	160	160
	Incubation Time (hr)	48	48	48	48	48	48	48	48
	Zeocin Conc. (uL/mL)	0	0	0	0	1	1	1	1
	Cells Concentrated (10 mL; +/-)	-	-	-	-	-	-	+	+
Expression Culture	YEP Media Volume (mL)	320	320	160	320	320	320	320	320
	KH ₂ PO ₄ Volume (mL)	40	40	20	40	40	40	40	40
	YNB Volume (mL)	40	40	20	40	40	40	40	40
	Biotin Volume (mL)	0.8	0.8	0.4	0.8	0.8	0.8	0.8	0.8
	Dextrose Volume (mL)	10	10	5	10	10	10	10	10
	Methanol Volume (mL)	1	1	5	1	1	1	1	1
	Inoculation Volume (mL)	5	5	5	10	5	5	10	10
	Incubation Temp (°C)	28	22	28	22	28	28	28	28
	Incubation RPM	180	160	200	160	160	160	160	160
	Incubation Time (hr)	96	96	72	96	96	96	96	96
	Daily Methanol Volume (mL)	2	2	5	5	2	2	2	2
	Flask Type	B	B	B	B	B	N	B	N
Attempts		0	1	5	1	1	1	1	1

+: Yes; -: No; B: Baffled; N: Normal.

5.2.3 Confirmation of recombinant protein expression

Purified TcCatF protein and expression media were run on sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) to confirm correct protein size and purification. SDS-PAGE and Western blot analyses were carried out according to Laemmli (1970) and Towbin *et al.* (1979). Briefly, TcCatF protein was separated by 12% SDS-PAGE in duplicate, one stained with Coomassie brilliant blue R stain for protein size visualisation, and the other transferred to low fluorescence PVDF (LF PVDF) membrane (0.45 µm pore size) using the Trans-Blot Turbo RTA Transfer Kit, LF PVDF (Bio-Rad) for Western blot. The LF PVDF membrane was washed with 10% TBS-T between each step. The membrane was blocked with 5% (w/v) dry skim milk in TBS-T incubated for 1 hour at room temperature, followed by incubation with mouse anti-His-tag antibody diluted 1:10,000 for 1 hour at room temperature with orbital shaking. Finally, reaction development was evaluated

using Clarity Western ECL Substrate Kit (Bio-Rad) as per manufacturer's instructions and imaged on the C-DiGit® Blot Scanner (LI-COR Biosciences, USA).

Mass spectrometry (MS/MS) was performed by La Trobe University's Comprehensive Proteomics Platform. Eluted TcCatF protein sample in liquid and SDS-PAGE gel form was dried using a SpeedVac Concentrator and Savant Refrigerated Vapor trap (ThermoFisher) before resolubilisation in 100 µL of solution of 8 M Urea, 100 mM Tris-HCl pH 8.3. One microlitre of 200 mM TCEP (tris(2-carboxyethyl)phosphine) was then added to the sample (and incubated overnight at 21 °C in a ThermoMixer (ThermoFisher)). Four microlitres of 1 M iodoacetamide (IAA; in water) was added the following day and the sample was incubated in the dark at 21 °C. 500 µL of 50 mM Tris-HCl (pH 8.3) and 1 µg of sequencing-grade trypsin (Promega, USA) were then added to samples and left overnight at 37 °C in an incubator. The digests were acidified with 1% (v/v) trifluoroacetic acid (TFA) and the peptides desalted on poly(styrene-divinylbenzene) copolymer (SDB) StageTips (3 M, USA) as described previously (Rappsilber *et al.*, 2007).

Trypsin-digested peptides were reconstituted in 0.1% (v/v) TFA and 2% (v/v) acetonitrile (ACN) and then loaded onto a guard column (C18 PepMap 100 µm inner diameter (ID) × 2 cm trapping column, Thermo-Fisher Scientific) at 5 µl/min and washed for 6 min before switching the guard column in line with the analytical column (Vydac MS C18, 3 µm, 300 Å and 75 µm ID × 25 cm). The separation of peptides was performed at 250 nl/min using a linear ACN gradient of buffer A (0.1% (v/v) formic acid, 2% (v/v) ACN) and buffer B (0.1% (v/v) formic acid, 80% (v/v) ACN), starting at 5% (v/v) buffer B to 30% in 65 min and then to 50% B at 78 min. The column was then eluted from peptides at 99% B for 5 min following equilibration at 5% B for 5 min. Data were collected on an Orbitrap Elite (ThermoFisher) in a data-dependent acquisition mode using m/z 300–1500 as MS scan range, collision-induced dissociation (CID) MS/MS spectra was collected for the 20 most intense ions. Dynamic exclusion parameters were set as described previously (Nguyen *et al.*, 2016). The Orbitrap Elite was operated in dual analyser mode, with the Orbitrap analyser being used for MS and the linear trap being used for MS/MS.

5.2.4 Protein activation analysis

pH activation of prev-TcCatF was attempted using citrate-phosphate buffer at pH 2.2, 3, 4, 5, 6, 7 and 8. Prev-TcCatF 2, an expression using method 2 (Table 5.4), was diluted 1:3 in citrate-phosphate buffer solution to the desired pH and was incubated at 37°C for 1, 2, 3, and 24 hours. Any change in protein was observed via SDS-PAGE.

5.3 Results

5.3.1 Characterisation of the full-length prev-TcCatF cDNA clone and predicted protein

The prev-TcCatF clone is 1,070 nucleotides long (Figure 5.2). Within the pPICZ α B + prev-TcCatF construct is a reading frame encoding a protein of 444 amino acids, of which 88 are predicted to comprise the α -factor signal peptide and 356 to comprise the propeptide sequence (predicted to begin at A₉₀). The mature protein is predicted to be 214 amino acids long, beginning at E₂₂₃ of the complete sequence specified by the construct (Figure 5.3).

TcCatF cDNA (1070 bp)

```

1      10      20      30      40      50      60
CTGCAGGACA ATACTCAGGA GGTGTAAAC CATTGACTGA ATTGAGAACG GATTTGATCG
70      80      90      100     110     120
ACAAGAAGAC CAAAGGTAGT ATCGAATTTG CCAGGCTTGG TCAACACATC AGTCCAAAAG
130     140     150     160     170     180
ACTTTGGTGC ATGGAATCAC TTCACCAGCT TCATTGAAAG GCACGATAAG GTCTACAGAA
190     200     210     220     230     240
ACGAGTCCGA AGCTCTGAAG AGGTTTGGGA TCTTTAAGAG AAATCTTGAG ATTATTAGAT
250     260     270     280     290     300
CTGCGCAAGA AAACGATAAG GGTACAGCTA TTTATGGTAT AAATCAGTTC GCTGACCTAT
310     320     330     340     350     360
CACCCGAGGA ATTTAAAAAA ACTCACTTGC CGCACACATG GAAGCAACCT GACCACCCAA
370     380     390     400     410     420
ACAGAATTGT GGAATTAGCC GCCGAAGGGG TTGATCCAAA AGAGCCACTG CCTGAATCGT
430     440     450     460     470     480
TCGATTGGAG AGAACATGGT GCCGTTACAA AAGTTAAAAC TGAAGGTCAT TGCGCAGCCT
490     500     510     520     530     540
GCTGGGCATT TTCTGTCACA GGAAATATTG AAGGACAATG GTTCTTGGCC AAGAAGAAAC
550     560     570     580     590     600
TTGTATCCTT GAGTGCTCAA CAGCTCTTGG ATTGTGATGT TGTGGATGAG GGTTGTAACG
610     620     630     640     650     660
GTGGATTTC TCTTGACGCT TATAAAGAAA TTGTTTCAAT GGGCGGCTTG GAACCAGAGG
670     680     690     700     710     720
ACAAGTATCC CTACGAAGCT AAGGCAGAGC AATGTAGATT AGTACCATCG GATATCGCTG
730     740     750     760     770     780
TTTATATTAA TGGATCAGTT GAGCTACCAC ATGATGAAGA GAAAATGAGA GCTTGGTTAG
790     800     810     820     830     840
TGAAGAAGGG GCCTATATCC ATTGGTATCA CCGTAGACGA TATACAGTTC TATAAAGGCG
850     860     870     880     890     900
GCGTTTCTCG TCCTACTACT TGTAGATTAT CTTCTATGAT TCATGGTGCT TTA CTGGTCG
910     920     930     940     950     960
GATACGGAGT CGAAAAAAAC ATCCCTTACT GGATTATTAA GAATTCTTGG GGACCTAATT
970     980     990     1,000   1,010   1,020
GGGGTGAGGA TGGATATTAC AGAATGGTGC GTGGAGAGAA CGCATGTCGC ATAAATAGAT
1,030   1,040   1,050   1,060   1,070
TTCCCACTTC AGCTGTTGTT CTACATCACC ATCATCATCA TTAATCTAGA

```

Figure 5.2: Codon-optimised, commercially synthesised TcCatF cDNA sequence used in the pPICZα B construct.

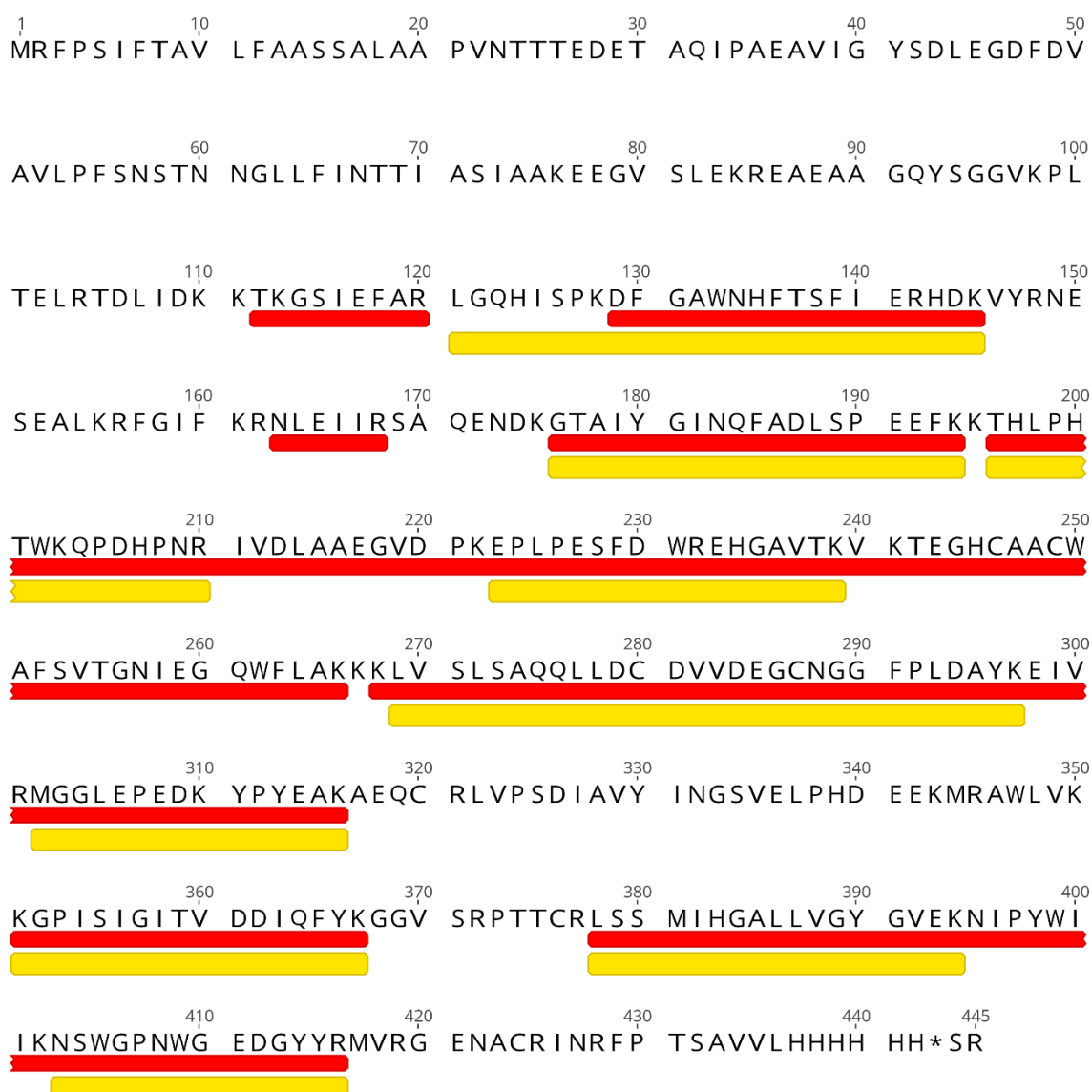


Figure 5.3: TcCatF and α -factor secretion signal amino acid sequence translated from pPICZ α B + TcCatF construct. Yellow: ms/ms fragments observed in prev-TcCatF 2 SDS-PAGE gel sample; red: ms/ms fragments observed in prev-TcCatF 2 liquid eluent sample; α -factor secretion signal: amino acids 1-88; mature protein: amino acids 223-442.

5.3.2 Expression of prev-TcCatF in *Pichia pastoris*

The DNA sequence encoding TcCatF and a hexahistidine tag at the C terminus was cloned into the yeast expression vector pPICZ α B. The recombinant plasmid was transformed into X33 *P. pastoris* for protein expression. Prev-TcCatF was expressed after methanol induction and purified from *P. pastoris* culture supernatants as shown by SDS-PAGE (Figure 5.4).

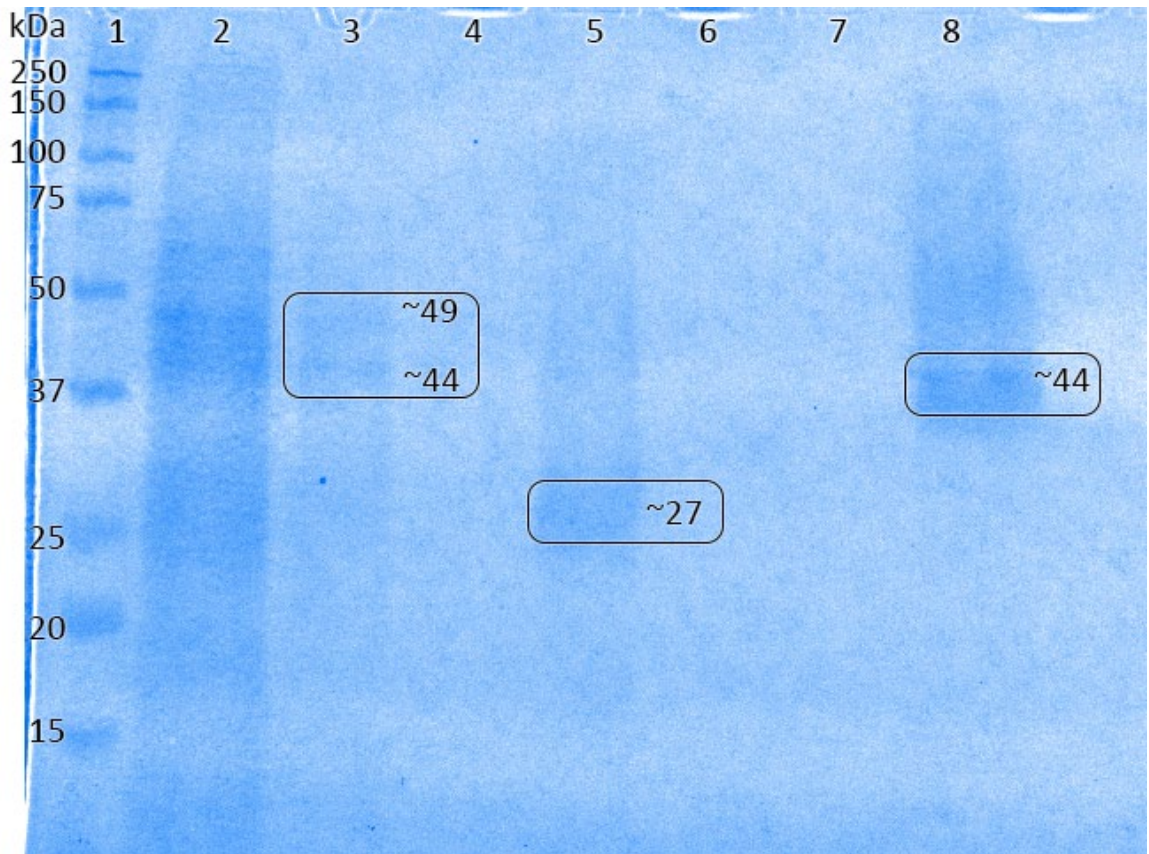


Figure 5.4: SDS-PAGE analysis of the expression culture supernatant using method 2. Lane 1: ladder; Lane 2: Ni-NTA column flow-through; Lane 3: Wash 1; Lane 4: Wash 2; Lane 5: prev-TcCatF 3 elution 1; Lane 6: prev-TcCatF 3 elution 2; Lane 7: prev-TcCatF 3 elution 3; Lane 8: prev-TcCatF 2 elution 1. Bands of interest correspond to the mature domain: 27 kDa; inactive precursor: 44 kDa; and whole sequence including an N-terminal α -factor secretion signal: 49 kDa. Faint bands of sizes ~49 kDa and ~44 kDa present in lane 3; faint band of size ~27 kDa in lane 5; faint band of size ~44 kDa in lane 8.

For purification of the recombinant protein, Ni-NTA resin was used to capture the hexahistidine-tagged protein in the yeast culture supernatant in a batch format, with an average yield of ~45 mg/L. The purified protein was analysed by SDS-PAGE which revealed a mobility of 27, 44, and 49 kDa, corresponding to the mature domain, the inactive precursor, and the whole sequence including an N-terminal α -factor secretion signal, respectively (Figure 5.11). Most purified protein was ~44 kDa in size. This result is consistent with the calculated molecular mass of a secreted protein containing the pro-mature region of TcCatF and a C-terminal hexahistidine tag.

Western Blot confirmed the presence of the hexahistidine tag by utilising mouse anti-His-tag antibody (data not shown), and mass spectrometry confirmed the protein sequence of the eluted prev-TcCatF protein corresponded to the TcCatF cDNA sequence (Figure 5.10).

5.3.3 Activation of prev-TcCatF

Incubation of prev-TcCatF with citrate-phosphate buffer had no effect on activity activation as measured by SDS-PAGE timelapse (Figure 5.5). No change in band size, indicating cleavage of pro-region, was observed.

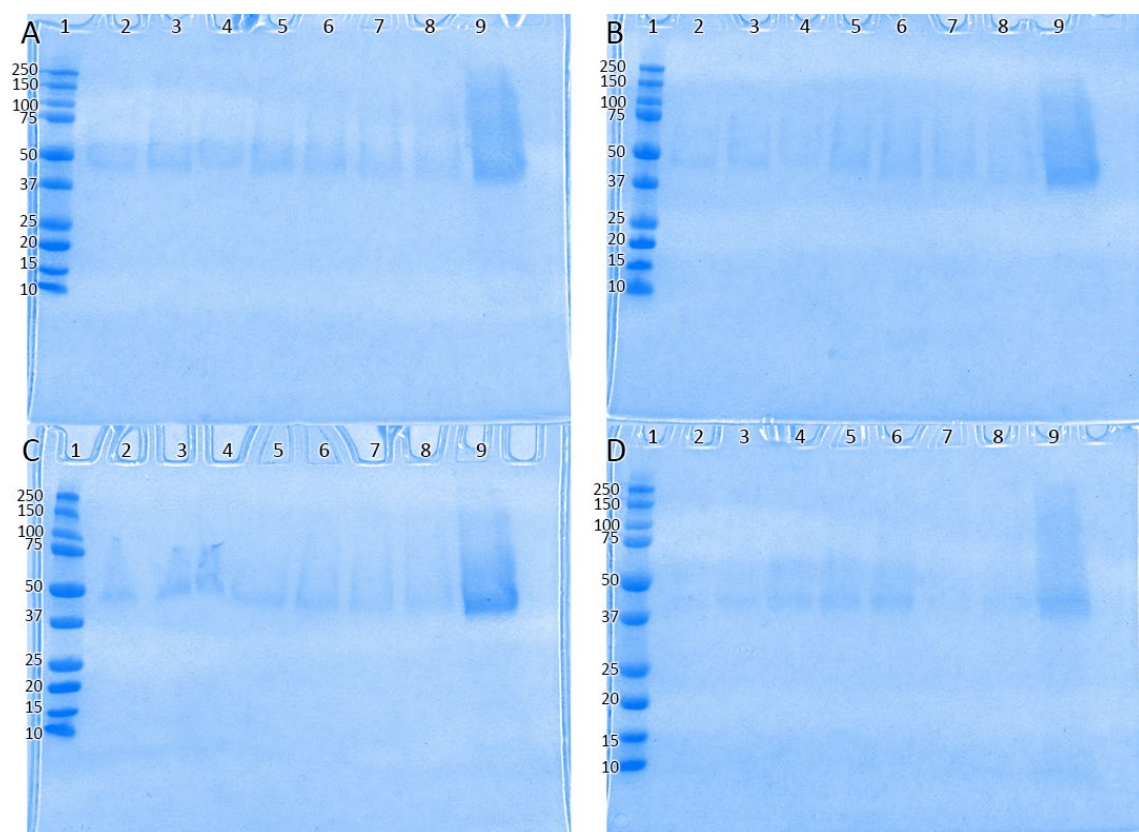


Figure 5.5: SDS-PAGE analysis of prev-TcCatF 2 purified protein at varying pH and incubation times. A: 1 hour incubation; B: 2 hours incubation; C: 3 hours incubation; D: 24 hours incubation.

Lane 1: ladder; Lane 2: prev-TcCatF 2 at pH 2.2; Lane 3: prev-TcCatF 2 at pH 3.0; Lane 4: prev-TcCatF 2 at pH 4.0; Lane 5: prev-TcCatF 2 at pH 5.0; Lane 6: prev-TcCatF 2 at pH 6.0; Lane 7: prev-TcCatF 2 at pH 7.0; Lane 8: prev-TcCatF 2 at pH 8.0; Lane 9: negative control prev-TcCatF 2 – no pH adjustment, no incubation time.

5.4 Discussion

This study was conceived to determine the function of *T. circumcincta* cathepsin F and its role in the survival of L4 in the abomasum of host sheep. Activity assays were planned once sufficient recombinant protein was produced. Additionally, the recombinant cathepsin F protein of this study was to be used in X-ray crystallography to determine the tertiary structure. Complications and difficulties arose, and due to time constraints and restricted laboratory access because of the Covid-19 pandemic, this study was abandoned. However, what has been done so far may act as a launching platform for the next researcher. This discussion will focus on what was done, what

could have been done better or differently, next steps, and what other researchers working on cathepsins have done in the past that could be applied here.

Cathepsin F of *T. circumcincta* has been shown to be the most abundant immunogenic protein secreted by L4 (Redmond *et al.*, 2006). The immune response against L4 influences parasite growth, mature size, and fecundity (Stear *et al.*, 1995a), but little is known about the targets of the immune response and how the parasite modulates immunity. Cathepsin L of *Fasciola hepatica*, a trematode species which causes severe liver damage, has been shown to cleave cattle IgG antibodies (Carmona *et al.*, 1993; A. M. Smith *et al.*, 1993). In this study we planned to explore whether cathepsin F of *T. circumcincta* would cleave sheep IgA antibodies. However, confirmation of this hypothesis requires further research.

Recombinant cathepsin F of *T. circumcincta* has previously been produced by Nisbet *et al.* (2013) for use in a multivalent vaccine against Teladorsagiosis. The protocol Nisbet *et al.* (2013) used was shared with our laboratory and a previous researcher attempted the method. The methods used in this study were standard kit manufacturer methods which have been successful in other cathepsin studies (Dağlioğlu, 2017), and as such were a good starting point. The prev-TcCatF produced by these standard manufacturer methods was deemed inadequate due to inactive protein, and a new attempt at cathepsin F recombinant protein cloning, transformation and expression was attempted with new-TcCatF.

Variables tested in this study included starter culture incubation RPM and Zeocin™ concentration, expression culture media volume (and its corresponding components), culture inoculation, incubation temperature, RPM and time, daily methanol volume, and flask type. Realistically, the incubation RPM, time and temperature differences, and the flask type differences should not affect the output to a great degree. What was most likely to affect the success of these expressions was the inclusion/exclusion of Zeocin™, and the volume of methanol added to the culture each day. Several expression protocols were conducted where Zeocin™ was omitted. Zeocin™ is an antibiotic for which the TcCatF *P. pastoris* constructs were resistant. Any contaminating bacteria would have been killed by the Zeocin™. If no Zeocin™ is present, it is possible that bacterial growth in the culture may occur and interfere with *P. pastoris* protein expression. Similarly, the volume of methanol added to the culture each day has a significant effect. In these constructs the AOX gene converts methanol into formaldehyde and energy for the cells to multiply and produce the protein of interest. It is known that no more than 0.5% total culture volume of methanol can be added each day as suggested in the *Pichia* Expression Kit manual, otherwise the methanol either overwhelms or kills off the *P. pastoris* cells and protein expression cannot occur (Mayson *et al.*, 2003).

This study had some success with prev-TcCatF expression, but the protein never appeared to have any activity. SDS-PAGE gels confirmed the correct protein size was being produced, though bands of several sizes were regularly present. Consistently, the inactive precursor sized protein was eluted. This would have been the ideal outcome because the risk of degradation of an active protein is avoided. Instead, a way to activate the protein would be required to test its function. In this instance when the eluted protein was pH adjusted, there appeared to be no change after several hours of observation. On several occurrences a protein of size ~27 kDa was shown on the SDS-PAGE gels, indicating a mature, or active, protein. A protein of size ~49 kDa was also shown on the SDS-PAGE gels several times. This size would correspond to the whole sequence including the secretion signal, however, the secretion signal should be cleaved during exocytosis. None of these eluted proteins showed any sign of activity.

Cysteine cathepsins require reducing and mildly acidic conditions for optimal activity, and it has been noted that all cathepsins, except cathepsin S, are irreversibly inactivated at pH 7 (Vidak *et al.*, 2019). This was unknown to us during prev-TcCatF expression, and the protocol used in this study adjusted the expression culture media to pH 7. It is quite possible that the TcCatF is being irreversibly inactivated as soon as it is secreted from the *P. pastoris* cells due to the pH of the supernatant. This could be why we do not see any activity from this protein. Maturation of cathepsins is complex, with pH-dependent autoactivation proposed to be the primary mechanism of activation for several cathepsin enzymes (Brömme *et al.*, 1993; Law *et al.*, 2003; Novinec *et al.*, 2012), while others have been shown to be activated during incubation with pepsin (Dağlioğlu, 2017; Linnevers *et al.*, 1997).

In other cathepsin studies (Brömme *et al.*, 1993; Dağlioğlu, 2017; Law *et al.*, 2003; Linnevers *et al.*, 1997; Novinec *et al.*, 2012; Puzer *et al.*, 2004; S. M. Smith *et al.*, 1989), fluorescent synthetic substrates were used to monitor protein activation. These are short peptides which bind to the active site and fluoresce for quantification, ultimately preventing further activity but allowing researchers to detect maturation and activity of the cathepsin. Further literature review should be undertaken to determine which substrate is most suitable for *T. circumcincta* cathepsin F and how to apply it. Many substrates exist for cathepsins as their active sites require slightly different amino acid sequences to bind. The methods in the studies mentioned above should be considered, as they have been shown to produce active protein.

A Western blot was performed to ensure the hexahistidine-tag was present on the protein. This his-tag binds the nickel in the Ni-NTA column for purification, and remaining protein passes through as waste. If there is a his-tag present then one can conclude it is the target protein that was captured by the resin. The Western blot confirmed that the protein being eluted was TcCatF, however, again multiple bands were observed at different sizes. Any other proteins potentially

being produced in the culture supernatant are unlikely to have a his-tag naturally and therefore are not going to bind the resin and be eluted with our TcCatF. These additional bands are most likely TcCatF but in the aforementioned conformational stages, possibly even as a dimer.

Mass spectrometry was used to confirm that the amino acid sequence of the produced prev-TcCatF matched with what was encoded in the construct. Based on the size determined in SDS-PAGE it was unlikely the sequence would differ dramatically, however, it is valuable to confirm that the sequence has not been altered during the transformation and cloning process, and to confirm whether the pro-region is still present in the eluted protein. Although the fragments did not cover the entire sequence of TcCatF, we can conclude that the correct sequence was produced, and the pro-region had not cleaved in the samples tested. This indicates a TcCatF with both the pro and mature domains.

The next step for confirmation of protein expression would be to utilise Circular Dichroism (CD) to confirm that the protein is being folded correctly. CD uses circularly polarised light to determine structural aspects of a sample (Fasman, 1996). In the case of TcCatF, it would give an approximation of alpha helices and beta sheets present in the protein. Comparison to homology modelling could determine whether the protein has been folded similarly or correctly. High expression of protein may overwhelm post-translational machinery of *P. pastoris* cells causing misfolding, or unprocessed or mis-localised protein (Cereghino *et al.*, 2000). The concentration of TcCatF calculated in this study does not indicate high expression, but CD may help to identify whether this is occurring. If prev-TcCatF had been shown to have folded correctly by CD, the next steps would have been to determine whether activation of TcCatF was prevented due to the eluent conditions or the expression method itself.

5.5 Conclusion

Although incomplete, this study has paved the way for future research into *T. circumcincta* cathepsin F protein expression. Expression of this protein is definitely possible; it just needs more time and experimentation for success. Producing an active cathepsin F protein would be greatly useful for understanding how *T. circumcincta* establishes within the host and may provide an interesting target for drug development and nematode control.

Chapter 6

General Discussion

6.1 Introduction

Mitigating the economic impact of infectious diseases in livestock animals, such as parasitic gastroenteritis caused by the parasitic nematode *Teladorsagia circumcincta*, will improve production output while enhancing the overall welfare and quality of life of affected animals. Internal parasites of sheep alone are estimated to cost \$436 million annually in Australia (Lane *et al.*, 2015), and even more worldwide, due to the costs associated with reduced production, disease control and treatment (Stear *et al.*, 2007).

This thesis describes a study on the secreted cathepsin F protein, the most abundant protein produced by fourth-stage larvae (L4; Redmond *et al.*, 2006), to determine whether this protein could be used as a target antigen for detection of resistance/susceptibility in sheep or initiating host immune responses to fight infection. To address these factors a thorough understanding of the protein itself is essential to provide insight for further research. This thesis aimed to describe the cathepsin F gene and gene structure, its protein structure, its relationship with cathepsins of close nematode relatives, and its functional role in nematode survival. To facilitate this, it was essential to determine a DNA extraction method suitable for individual nematode specimens. The selected method was then used in conjunction with Oxford Nanopore Technology to sequence the *T. circumcincta* genome for analysis of the cathepsin F gene. Additionally, the production of recombinant cathepsin F protein was necessary for functional analyses.

6.2 Cathepsin F bioinformatic analysis

Many cysteine proteases have been studied to identify conserved DNA and amino acid sequence motifs across species, to determine functional aspects, and confirm protein structure. As a result, parasitic cysteine proteases have been shown to be involved in parasite stage transition, invasion of host tissues, nutrient uptake, and immune evasion (Barrett *et al.*, 1981; Carmona *et al.*, 1993; Dana *et al.*, 2020; Somoza *et al.*, 2002; Turk *et al.*, 2012; Vidak *et al.*, 2019). Cathepsin F has been identified in several species, all with slight variation in sequence and protein structure but with conserved familial motifs (Deussing *et al.*, 2000; J. M. Kang *et al.*, 2013; T. H. Kang *et al.*, 2004; Redmond *et al.*, 2006; Vidak *et al.*, 2019; Wex *et al.*, 1999). Some parasite proteins have evolved to adapt pre-existing roles to immune-modulatory roles (Maizels *et al.*, 2018), while other proteins have convergently evolved structures to interact with host receptors (Johnston *et al.*, 2017). Bioinformatic analyses allow researchers to identify the potential roles and functions of proteins as well as help determine their history or origin.

Chapter 2 used a bioinformatic approach to investigate *T. circumcincta* cathepsin F. Specifically, a putative gene encoding the protein was assembled, the gene structure was described, several potential polymorphisms were discovered, close homologues were found to be absent, and the structure of the protein was predicted. Previous research into *T. circumcincta* cathepsin F is limited

to a handful of studies. A previous study investigating *T. circumcincta* cathepsin F did basic annotation of the protein sequence, theoretical molecular weight estimation, phylogenetic analysis, and confirmation of glycosylation (Redmond *et al.*, 2006). The work in Chapter 2 found cathepsin F is a relatively recently evolved cysteine protease which does not fall clearly into either of the cathepsin L or F subfamilies. Homology modelling of the protein structure indicated that *T. circumcincta* cathepsin F lacks a cystatin-like domain making it structurally similar to a typical cathepsin L. Additionally, crucial amino acid motifs which characterise cathepsins into their different families were not distinctly cathepsin F or cathepsin L. As a result the molecule appeared to be a hybrid, and exploration of protein homology in related nematode species was conducted (Sloan *et al.*, 2020).

Cathepsin F of *T. circumcincta* was not similar to proteins of closely related nematodes of the same subfamily and lacked apparent orthologues. The absence of close orthologues indicates *T. circumcincta* cathepsin F may have evolved relatively recently. It is possible that during the adaptation of *T. circumcincta* to parasitism, several genes were rapidly altered and selected to manage host immune responses. There may be other newly evolved genes, perhaps immunosuppressive, that are yet to be identified that could benefit from bioinformatic analysis as was done in Chapter 2. Identifying immunomodulatory genes and determining their function and role could greatly enhance our ability to control this parasite, to ensure a high level of animal welfare and production output.

Cathepsin F of *T. circumcincta* had not been explored in detail bioinformatically prior to this thesis, and this has provided a good foundation for further work. For example, confirmation of bioinformatic hypotheses requires further experimental data through examination of the *T. circumcincta* genome, or through functional assays using secreted or recombinantly produced proteins.

6.3 Individual nematode DNA extraction

Analysis of variation in the cathepsin F gene required a DNA extraction method for individual nematodes. Previous studies have used a variety of methods to extract DNA from nematodes, either pooled or individually (Choi *et al.*, 2017; Palevich *et al.*, 2019; Parkinson *et al.*, 2004). However, a comparison of DNA extraction methods to determine the optimal method for DNA extraction from *T. circumcincta* nematodes had not previously been conducted. Although previous studies have utilized differing methods of DNA extraction for *T. circumcincta*, no consensus has been reached regarding an optimal method.

The aim of Chapter 3 was to develop a reliable DNA extraction protocol to produce high quality DNA for genome sequencing and phylogenetic analyses. 11 extraction kits were compared based

on DNA quality, yield, and processing time, and included chelating, precipitation, and silica-binding methods. NanoDrop 2000™ spectrophotometry and Qubit™ fluorometry were used to determine quality and concentration of DNA extracted, and 4 of the extraction methods were deemed capable of producing appropriate quality and concentration levels.

Larval cuticles have been shown to negatively affect DNA extraction because the cuticle interferes with chemical lysis of cells. Chemical lysis is preferred over physical lysis for extraction, because of the retention of intact and long DNA strands (Seesao *et al.*, 2014) for long-read sequencing. Larval exsheathment is sometimes considered necessary for nematodes to access suitable tissue, however, the extra processing can damage the cells and depending on the specimen size can detrimentally affect both DNA quality and yield. Chapter 3 confirmed that for *T. circumcincta* larval exsheathment negatively impacted both concentration and purity of extracted DNA. Of the 11 protocols compared, a *Schistosoma sp.* DNA extraction method (Schi) was most suitable for individual *T. circumcincta* nematode DNA extraction due to its resulting DNA concentration, purity, and relatively fast processing time.

Determination of an appropriate DNA extraction method enabled DNA sequencing of individual nematode specimens for further cathepsin F genetic analysis. As technology advances and methods change and update, further comparative analyses for single nematode DNA extraction will be required. More methods than those attempted in Chapter 3 could be compared, and the methods themselves could be further refined for optimal output. Additionally, exploring whether bleach could be used as an additional cleansing step may be worthwhile. Extraction of DNA from individual specimens is particularly valuable for research utilising PCR.

6.4 MinION™ sequencing of *T. circumcincta*

Oxford Nanopore Technology (ONT) sequencing is a relatively new technology, with no previous reports of this technology being utilised for *T. circumcincta* and was conducted following the Schi DNA extraction method determined in Chapter 3 (Sloan *et al.*, 2021). Previous genome sequencing has used PacBio for long-read sequencing, and Illumina and Genome Sequencer Titanium FLX for short-read sequencing (Choi *et al.*, 2017; Palevich *et al.*, 2019). Chapter 2 found that the gene encoding cathepsin F in the Choi *et al.* (2017) draft *T. circumcincta* genome was incomplete and misassembled, and contained many gaps (Sloan *et al.*, 2020). As such, a reassembly was required to look more specifically at the cathepsin F gene. A complete gene was not able to be constructed with the data available at the time, and further sequencing was determined necessary.

Chapter 4 generated over 3 million reads using ONT's MinION™ long-read sequencer, and these reads along with PacBio reads sequenced by the Wellcome Trust Sanger Institute were used to assemble a draft *T. circumcincta* genome of ~114.3 Mbp. This draft consisted of 9,219 contigs with

35.45X coverage, a minimum and maximum sequence length of 1,074 and 176,649 bp, respectively, and a GC content of 43.7%. A total of 28,588 protein-coding genes were predicted. This contrasts with the Choi *et al.* (2017) draft genome of ~701 Mbp size, 81,730 contigs, and an estimated 25,532 protein-coding genes. Their genome was constructed with short-read sequences and DNA extracted from pooled, partially in-bred drug-susceptible strains of *T. circumcincta* for the purpose of comparing genome-wide single nucleotide and copy number variants of drug-resistant *T. circumcincta* strains. The Choi *et al.* (2017) draft calculated a genome completeness of 98% using the Core Eukaryotic Genes Mapping Approach (CEGMA, Parra *et al.*, 2007), but with further refinement, annotation, and the inclusion of long-read sequencing data it is likely the genome size may reduce closer to 58.6 Mbp, a *T. circumcincta* complete genome size estimated by Leroy *et al.* (2003) using flow cytometry. It is probable that a substantial portion of the Choi *et al.* (2017) draft genome as it stands consists of duplicates and repeat regions.

The analysis completed in Chapter 2 saw many fragmented and partial duplicates of the cathepsin F gene in the Choi *et al.* (2017) draft genome. Should the draft genome be annotated completely and polished, it is likely to become more condensed. The cathepsin F gene is unlikely to be alone in this regard, and other genes are likely also fragmented or duplicated, and require further detailed analysis. This thesis provides a framework for how that analysis can be achieved. Comparing specific genes in the draft genome of Chapter 4 and the Choi *et al.* (2017) draft will not only help with the overall polish of the genome but will also provide insights into the workings of *T. circumcincta* itself. Using programs such as CEGMA to ensure conserved genes are maintained while annotating predicted genes, as well as identifying and condensing accidental repeat regions will greatly improve the quality of the genome assemblies. Additionally, the more long-read sequencing data that can be included, the more likely that quality will be achieved.

6.5 Cathepsin F gene

The two putative alleles of cathepsin F identified in Chapter 4 were each found on separate contigs; contig00000245 (tig245) and contig00000282 (tig282). Tig245 was 59,445 bp in length, made up of 216 reads, and had a GC content of 45.9%. The complete *T. circumcincta* cathepsin F gene present on this contig had a coding length of 1,095 bp over 10 exons, and a total length of 11,564 bp. Tig282 was 34,165 bp in length, made up of 73 reads, and had a GC content of 45.9%. The complete *T. circumcincta* cathepsin F gene present on this contig had a coding length of 1,101 bp over 10 exons, and a total length of 9,663 bp. Both contigs also contained additional partial genes. Considering these two complete genes are likely to be alleles rather than distinct genes, as discussed in Chapter 4, these partial genes may represent additional alleles because the flanking predicted genes matched across contigs. The partial genes on tig245 contained many substitutions and because the draft is made up of several different nematode specimens, could be indicative of

high variation in the gene amongst individuals, each with a unique diploid genome. The incomplete gene on tig282, however, matched very closely with both complete genes on tig245 and tig282, and may be a repeat error.

Substitutions which differentiated the genes on tig245 and tig282 from Tci-CF-1 were also observed and had been previously identified in Chapter 2 (Sloan *et al.*, 2020), indicating this gene has several polymorphisms and there may be great variation between individual nematode specimens and populations.

6.6 *T. circumcincta* microbiome analysis

It is crucially important in genome assembly that the data used for construction belongs solely to the organism of interest. Gastrointestinal worms are challenging when the organism cannot be grown *in vitro* (Luque *et al.*, 2010). Contaminating DNA from other gastrointestinal organisms and the host need to be filtered out before assembly. Whether that is achieved prior to DNA extraction with additional cleansing steps such as with bleach, or completed post-sequencing depends on the organism.

Chapter 3 confirmed that the Schi method was able to extract *T. circumcincta* DNA at a high quality, while Chapter 4 further highlighted that the DNA extracted contained a multitude of sequences from other species. Additionally, due to the lack of host DNA in the extracted samples, it can be concluded thorough washing was able to minimize external contaminants (Sloan *et al.*, 2021), but the physiology of the nematode allows for the presence of other species within its own gastrointestinal tract. Resultantly, post-sequencing filtering of sequences is required to ensure an assembly uses only *T. circumcincta* DNA.

Most other sequences were classified as *E. coli*, with additional bacterial and fungal species also present. Of particular interest was the presence of Orbiliomycetes, a nematophagous fungus that is common in temperate regions, where *T. circumcincta* is most prominent (Yang *et al.*, 2007). These Orbiliaceae have evolved specialised mechanisms to trap and digest nematodes. All other species present were either known to be common in soil, vegetation, or the gastrointestinal tracts of animals, and as such their presence was expected. It would be beneficial in future to sequence an additional sample of the environment the nematode specimens were in to conduct a differential analysis to identify the microbes specifically enriched within the nematode versus the host.

Other nematode species also identified included *Haemonchus contortus*, *Ostertagia* spp., and *Chabaudstrongylus ninhae*, among others. It is likely that these nematode species have genomic sequences that are similar to *T. circumcincta* and may have been misassigned due to the incompleteness of the *T. circumcincta* genome. Alternatively, there may be errors in the sequence.

Mammalian species such as *Ovis aries* and *Odocoileus virginianus* (white-tailed deer) were also observed. Similarly, these DNA sequences likely belong to sheep, the host, but share similarities with deer, and incomplete sequences or sequencing error has resulted in incorrect classification.

6.7 Cathepsin F protein expression

The functional role of cathepsin F has not yet been determined, despite its use in a multivalent vaccine and being a known target for IgA antibodies in sheep (Nisbet *et al.*, 2013). Cathepsin F of *T. circumcincta* has been shown to be the most abundant immunogenic protein secreted by L4 (Redmond *et al.*, 2006), and the immune response against L4 influences parasite growth, mature size, and fecundity (Stear *et al.*, 1995a). Little is known about the targets of the immune response and how the parasite modulates immunity.

The onset of the Covid-19 pandemic significantly impacted the progress of the work in Chapter 5, ultimately leading to its abandonment. However, the aim for the research was to successfully produce recombinant protein and determine the cathepsin F function. This study had some success with recombinant cathepsin F expression, but activation of the protein and any subsequent activity failed. Whether this inactive protein can be useful as a target for ELISA diagnostic tests or in a vaccine is unclear and has not yet been explored. Inactive protein may be useful if correct and successful protein folding has occurred and can be confirmed, as the active site cleft will be blocked but the overall shape of the protein will be correct and may bind to other molecules.

Once an active recombinant protein is produced, it would be valuable to test whether cathepsin F of *T. circumcincta* digests IgA in the way cathepsin L of *Fasciola hepatica* cleaves host cattle IgG (A. M. Smith *et al.*, 1993), or if it has any effect on other common proteins within the host gut. Further, once activity of this protein is determined, it will become quite important to investigate the relevance of the polymorphisms. Are some variants more efficacious, or faster acting, or do different variants target different molecules? Valuable information for nematode control will be garnered from these answers.

Although incomplete, this study has paved the way for future research into *T. circumcincta* cathepsin F protein expression. This thesis indicates that expression of *T. circumcincta* cathepsin F is feasible, and that a well-planned experiment and time for such an experiment to be completed would ensure success. Producing an active cathepsin F protein would be of great value towards understanding how *T. circumcincta* establishes within the host and may provide an interesting target for drug development and nematode control.

6.8 Conclusion

As with most scientific studies, the answers presented in this thesis provoke more questions for further research. Most importantly, what is the function of cathepsin F, how does it contribute to nematode establishment in the host, and how does the polymorphism variation impact function and immune evasion?

While this thesis was in preparation, minimal research in this area was performed by other research institutions despite the importance of *T. circumcincta* for agricultural production. Additionally, no research was published on cathepsin F by other research institutions during the same time period. This project demonstrated that a targeted and detailed analysis of proteins can be of great benefit to the overall control of this parasite. By truly understanding the functions, roles and genetics of selected proteins, a targeted effort for specific outcomes using treatments and control methods can be achieved. The results presented in this thesis have improved our understanding of cathepsin F, particularly the presence of polymorphisms in the gene which may influence how successfully this nematode will evade the host immune system.

In conclusion, the use of a multidisciplinary approach including bioinformatics, genomics, animal science and parasitology has enhanced our understanding of a previously enigmatic but key protein in a neglected but important parasite. This protein has recently evolved and may represent an adaptation to parasitism that is unique to *T. circumcincta*. Genome sequencing using ultra long reads has improved the genome assembly and identified polymorphisms. Further, the improved genome assembly will allow bioinformaticians and parasitologists to identify additional polymorphisms in other genes to further research towards the control of *T. circumcincta*.

References

- Abate, A. R., Hung, T., Sperling, R. A., Mary, P., Rotem, A., Agresti, J. J., . . . Weitz, D. A. (2013). DNA sequence analysis with droplet-based microfluidics. *Lab Chip*, 13(24), 4864-4869. doi:10.1039/c3lc50905b
- Abongwa, M., Martin, R. J., & Robertson, A. P. (2017). A Brief Review on the Mode of Action of Antinematodal Drugs. *Acta Vet (Beogr)*, 67(2), 137-152. doi:10.1515/acve-2017-0013
- Adewale, B. A. (2020). Will long-read sequencing technologies replace short-read sequencing technologies in the next 10 years? *Afr J Lab Med*, 9(1), 1340. doi:10.4102/ajlm.v9i1.1340
- Alexopoulos, C. J., Mims, C. W., & Blackwell, M. (1996). *Introductory Mycology* (3rd ed.). New York, USA: John Wiley & Sons.
- Amarante, A. F., Susin, I., Rocha, R. A., Silva, M. B., Mendes, C. Q., & Pires, A. V. (2009). Resistance of Santa Ines and crossbred ewes to naturally acquired gastrointestinal nematode infections. *Vet Parasitol*, 165(3-4), 273-280. doi:10.1016/j.vetpar.2009.07.009
- Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biol*, 21(1), 30. doi:10.1186/s13059-020-1935-5
- Ashrafi, K., Sharifdini, M., Heidari, Z., Rahmati, B., & Kia, E. B. (2020). Zoonotic transmission of *Teladorsagia circumcincta* and *Trichostrongylus* species in Guilan province, northern Iran: molecular and morphological characterizations. *BMC Infect Dis*, 20(1), 28. doi:10.1186/s12879-020-4762-0
- Bahirathan, M., Miller, J. E., Barras, S. R., & Kearney, M. T. (1996). Susceptibility of Suffolk and Gulf Coast Native suckling lambs to naturally acquired strongylate nematode infection. *Vet Parasitol*, 65(3-4), 259-268. doi:10.1016/s0304-4017(96)00969-7
- Bain, R. K. (1999). Irradiated vaccines for helminth control in livestock. *Int J Parasitol*, 29(1), 185-191. doi:10.1016/s0020-7519(98)00187-8
- Barrett, A. J., & Kirschke, H. (1981). Cathepsin B, Cathepsin H, and cathepsin L. *Meth Enzymol*, 80 Pt C, 535-561. doi:10.1016/s0076-6879(81)80043-2
- Barrett, A. J., & Rawlings, N. D. (2001). Evolutionary lines of cysteine peptidases. *Biol Chem*, 382(5), 727. doi:10.1515/BC.2001.088
- Bartley, D. J., Jackson, E., Johnston, K., Coop, R. L., Mitchell, G. B., Sales, J., & Jackson, F. (2003). A survey of anthelmintic resistant nematode parasites in Scottish sheep flocks. *Vet Parasitol*, 117(1-2), 61-71. doi:10.1016/j.vetpar.2003.07.023
- Bauer, R., Oberwinkler, F., & Vánky, K. (1997). Ultrastructural markers and systematics in smut fungi and allied taxa. *Can J Bot*, 75(8), 1273-1314. doi:10.1139/b97-842

- Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Arch Dis Child Educ Pract Ed*, 98(6), 236-238. doi:10.1136/archdischild-2013-304340
- Bender, W., Spierer, P., & Hogness, D. S. (1983). Chromosomal walking and jumping to isolate DNA from the Ace and rosy loci and the bithorax complex in *Drosophila melanogaster*. *J Mol Biol*, 168(1), 17-33. doi:10.1016/s0022-2836(83)80320-9
- Bentley, D. R. (2006). Whole-genome re-sequencing. *Curr Opin Genet Dev*, 16(6), 545-552. doi:10.1016/j.gde.2006.10.009
- Berger, L., Speare, R., & Hyatt, A. (1999). Chytrid fungi and amphibian declines: overview, implications and future directions. In *Declines and disappearances of Australian frogs*. Canberra, Australia: Environment Australia.
- Bishop, S. C., Bairden, K., McKellar, Q. A., Park, M., & Stear, M. J. (1996). Genetic parameters for faecal egg count following mixed, natural, predominantly *Ostertagia circumcincta* infection and relationships with live weight in young lambs. *Anim Sci*, 63(3), 423-428. doi:10.1017/S1357729800015319
- Bisset, S. A., Vlassoff, A., Morris, C. A., Southey, B. R., Baker, R. L., & Parker, A. G. H. (2011). Heritability of and genetic correlations among faecal egg counts and productivity traits in Romney sheep. *New Zealand J Agric Res*, 35(1), 51-58. doi:10.1080/00288233.1992.10417701
- Blaxter, M., & Koutsovoulos, G. (2015). The evolution of parasitism in Nematoda. *Parasitology*, 142 Suppl 1(Suppl 1), S26-39. doi:10.1017/S0031182014000791
- Bonnet, C., Volat, B., Bardin, R., Degranges, V., & Montuelle, B. (1997). Use of immunofluorescence technique for studying a *Nitrobacter* population from wastewater treatment plant following discharge in river sediments: First experimental data. *Water Res*, 31(3), 661-664. doi:10.1016/S0043-1354(96)00094-2
- Borgsteede, F., Verkaik, J., Moll, L., Dercksen, D., Vellema, P., & Bavinck, G. (2010). [How widespread is resistance to ivermectin among gastrointestinal nematodes in sheep in The Netherlands?]. *Tijdschr Diergeneesk*, 135(21), 782-785. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/21141150>
- Bott, N. J., Campbell, B. E., Beveridge, I., Chilton, N. B., Rees, D., Hunt, P. W., & Gasser, R. B. (2009). A combined microscopic-molecular method for the diagnosis of strongylid infections in sheep. *Int J Parasitol*, 39(11), 1277-1287. doi:10.1016/j.ijpara.2009.03.002
- Brenner, D. J., Krieg, N. R., & Staley, J. T. (2005). *Bergey's Manual® of Systematic Bacteriology Volume Two: The Proteobacteria (Part C)* (G. Garrity, D. J. Brenner, N. R. Krieg, & J. T. Staley Eds. Vol. 2). Boston, USA: Springer US.
- Brömme, D., Bonneau, P. R., Lachance, P., Wiederanders, B., Kirschke, H., Peters, C., . . . Vernet, T. (1993). Functional expression of human cathepsin S in *Saccharomyces cerevisiae*.

- Purification and characterization of the recombinant enzyme. *J Biol Chem*, 268(7), 4832-4838. doi:10.1016/S0021-9258(18)53472-4
- Burgess, C. G., Bartley, Y., Redman, E., Skuce, P. J., Nath, M., Whitelaw, F., . . . Jackson, F. (2012). A survey of the trichostrongylid nematode species present on UK sheep farms and associated anthelmintic control practices. *Vet Parasitol*, 189(2-4), 299-307. doi:10.1016/j.vetpar.2012.04.009
- Cafrune, M. M., Aguirre, D. H., & Rickard, L. G. (2001). First report of *Lamanema chavez* (Nematoda: Trichostrongyloidea) in llamas (*Lama glama*) from Argentina. *Vet Parasitol*, 97(2), 165-168. doi:10.1016/s0304-4017(01)00379-x
- Callaway, T. R., Dowd, S. E., Edrington, T. S., Anderson, R. C., Krueger, N., Bauer, N., . . . Nisbet, D. J. (2010). Evaluation of bacterial diversity in the rumen and feces of cattle fed different levels of dried distillers grains plus solubles using bacterial tag-encoded FLX amplicon pyrosequencing1. *J Anim Sci*, 88(12), 3977-3983. doi:10.2527/jas.2010-2900
- Capece, B. P., Virkel, G. L., & Lanusse, C. E. (2009). Enantiomeric behaviour of albendazole and fenbendazole sulfoxides in domestic animals: pharmacological implications. *Vet J*, 181(3), 241-250. doi:10.1016/j.tvjl.2008.11.010
- Carmona, C., Dowd, A. J., Smith, A. M., & Dalton, J. P. (1993). Cathepsin L proteinase secreted by *Fasciola hepatica* in vitro prevents antibody-mediated eosinophil attachment to newly excysted juveniles. *Mol Biochem Parasitol*, 62(1), 9-17. doi:10.1016/0166-6851(93)90172-t
- Cereghino, J. L., & Cregg, J. M. (2000). Heterologous protein expression in the methylotrophic yeast *Pichia pastoris*. *FEMS Microbiol Rev*, 24(1), 45-66. doi:10.1111/j.1574-6976.2000.tb00532.x
- Cernanska, D., Varady, M., & Corba, J. (2006). A survey on anthelmintic resistance in nematode parasites of sheep in the Slovak Republic. *Vet Parasitol*, 135(1), 39-45. doi:10.1016/j.vetpar.2005.09.001
- Chaucheyras-Durand, F., & Ossa, F. (2014). REVIEW: The rumen microbiome: Composition, abundance, diversity, and new investigative tools. *Prof Anim Sci*, 30(1), 1-12. doi:10.15232/s1080-7446(15)30076-0
- Chen, R. (2012). Bacterial expression systems for recombinant protein production: *E. coli* and beyond. *Biotechnol Adv*, 30(5), 1102-1107. doi:10.1016/j.biotechadv.2011.09.013
- Chmielecki, J., & Meyerson, M. (2014). DNA sequencing of cancer: what have we learned? *Annu Rev Med*, 65(1), 63-79. doi:10.1146/annurev-med-060712-200152
- Choi, Y. J., Bisset, S. A., Doyle, S. R., Hallsworth-Pepin, K., Martin, J., Grant, W. N., & Mitreva, M. (2017). Genomic introgression mapping of field-derived multiple-anthelmintic resistance

- in *Teladorsagia circumcincta*. *PLoS Genet*, 13(6), e1006857.
doi:10.1371/journal.pgen.1006857
- Coltman, D. W., Wilson, K., Pilkington, J. G., Stear, M. J., & Pemberton, J. M. (2001). A microsatellite polymorphism in the gamma interferon gene is associated with resistance to gastrointestinal nematodes in a naturally-parasitized population of Soay sheep. *Parasitology*, 122(Pt 5), 571-582. doi:10.1017/s0031182001007570
- Coop, R. L., Huntley, J. F., & Smith, W. D. (1995). Effect of dietary protein supplementation on the development of immunity to *Ostertagia circumcincta* in growing lambs. *Res Vet Sci*, 59(1), 24-29. doi:10.1016/0034-5288(95)90025-x
- Cydzik-Kwiatkowska, A., & Zielinska, M. (2016). Bacterial communities in full-scale wastewater treatment systems. *World J Microbiol Biotechnol*, 32(4), 66. doi:10.1007/s11274-016-2012-9
- Dağlıoğlu, C. (2017). Cloning, expression, and activity analysis of human cathepsin C in the yeast *Pichia pastoris*. *Turk J Biol*, 41(5), 746-753. doi:10.3906/biy-1704-4
- Dana, D., & Pathak, S. K. (2020). A Review of Small Molecule Inhibitors and Functional Probes of Human Cathepsin L. *Molecules*, 25(3), 698. doi:10.3390/molecules25030698
- Delbrassinne, L., & Mahillon, J. (2016). Bacillus: Occurrence. In B. Caballero, P. M. Finglas, & F. Toldrá (Eds.), *Encyclopedia of Food and Health* (pp. 307-311). Oxford, UK: Academic Press.
- Deussing, J., Tisljar, K., Papazoglou, A., & Peters, C. (2000). Mouse cathepsin F: cDNA cloning, genomic organization and chromosomal assignment of the gene. *Gene*, 251(2), 165-173. doi:10.1016/s0378-1119(00)00196-7
- Didier, E. S. (2005). Microsporidiosis: an emerging and opportunistic infection in humans and animals. *Acta Trop*, 94(1), 61-76. doi:10.1016/j.actatropica.2005.01.010
- Dominguez Del Angel, V., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Vinnere Pettersson, O., . . . Lantz, H. (2018). Ten steps to get started in Genome Assembly and Annotation. *F1000Res*, 7. doi:10.12688/f1000research.13598.1
- Doyle, S. R., Sankaranarayanan, G., Allan, F., Berger, D., Jimenez Castro, P. D., Collins, J. B., . . . Holroyd, N. (2019). Evaluation of DNA Extraction Methods on Individual Helminth Egg and Larval Stages for Whole-Genome Sequencing. *Front Genet*, 10, 826. doi:10.3389/fgene.2019.00826
- Duffy, M. S., Cevasco, D. K., Zarlenga, D. S., Sukhumavasi, W., & Appleton, J. A. (2006). Cathepsin B homologue at the interface between a parasitic nematode and its intermediate host. *Infect Immun*, 74(2), 1297-1304. doi:10.1128/IAI.74.2.1297-1304.2006

- Echevarria, F., Borba, M. F., Pinheiro, A. C., Waller, P. J., & Hansen, J. W. (1996). The prevalence of anthelmintic resistance in nematode parasites of sheep in southern Latin America: Brazil. *Vet Parasitol*, 62(3-4), 199-206. doi:10.1016/0304-4017(95)00906-x
- Eddi, C., Caracostantogolo, J., Pena, M., Schapiro, J., Marangunich, L., Waller, P. J., & Hansen, J. W. (1996). The prevalence of anthelmintic resistance in nematode parasites of sheep in southern Latin America: Argentina. *Vet Parasitol*, 62(3-4), 189-197. doi:10.1016/0304-4017(95)00905-1
- Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., . . . Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910), 133-138. doi:10.1126/science.1162986
- Elkins, K. M. (2012). *Forensic DNA Biology : A Laboratory Manual*. San Diego, USA: Elsevier Science & Technology.
- Elseadawy, R., Abbas, I., Al-Araby, M., Hildreth, M. B., & Abu-Elwafa, S. (2019). First Evidence of *Teladorsagia circumcincta* Infection in Sheep from Egypt. *J Parasitol*, 105(4), 484-490. doi:10.1645/18-202
- Farm Health First. (2021). Diseases & Solutions: Gastrointestinal Worms in Sheep. Retrieved from <https://www.farmhealthfirst.com/disease-solution/gastrointestinal-worms-in-sheep>
- Fasman, G. D. (1996). *Circular Dichroism and the Conformational Analysis of Biomolecules*. New York, USA: Springer US.
- Gamble, H. R., & Zajac, A. M. (1992). Resistance of St. Croix lambs to *Haemonchus contortus* in experimentally and naturally acquired infections. *Vet Parasitol*, 41(3-4), 211-225. doi:10.1016/0304-4017(92)90081-j
- Gasser, R. B., Chilton, N. B., Hoste, H., & Beveridge, I. (1993). Rapid sequencing of rDNA from single worms and eggs of parasitic helminths. *Nucleic Acids Res*, 21(10), 2525-2526. doi:10.1093/nar/21.10.2525
- Gonzalez, J. F., Hernandez, J. N., Machin, C., Perez-Hernandez, T., Wright, H. W., Corripio-Miyar, Y., . . . Nisbet, A. J. (2019). Impacts of breed type and vaccination on *Teladorsagia circumcincta* infection in native sheep in Gran Canaria. *Vet Res*, 50(1), 29. doi:10.1186/s13567-019-0646-y
- Good, B., Hanrahan, J. P., Crowley, B. A., & Mulcahy, G. (2006). Texel sheep are more resistant to natural nematode challenge than Suffolk sheep based on faecal egg count and nematode burden. *Vet Parasitol*, 136(3-4), 317-327. doi:10.1016/j.vetpar.2005.12.001
- Grillo, V., Jackson, F., Cabaret, J., & Gilleard, J. S. (2007). Population genetic analysis of the ovine parasitic nematode *Teladorsagia circumcincta* and evidence for a cryptic species. *Int J Parasitol*, 37(3-4), 435-447. doi:10.1016/j.ijpara.2006.11.014

- Groves, M. R., Taylor, M. A. J., Scott, M., Cummings, N. J., Pickersgill, R. W., & Jenkins, J. A. (1996). The prosequence of procarricain forms an α -helical domain that prevents access to the substrate-binding cleft. *Structure*, 4(10), 1193-1203. doi:10.1016/s0969-2126(96)00127-x
- Gruner, L., Bouix, J., Vu Tien Khang, J., Mandonnet, N., Eychenne, F., Cortet, J., . . . Limouzin, C. (2004). A short-term divergent selection for resistance to *Teladorsagia circumcincta* in Romanov sheep using natural or artificial challenge. *Genet Sel Evol*, 36(2), 217-242. doi:10.1186/1297-9686-36-2-217
- Gupte, M., Kulkarni, P., & Ganguli, B. N. (2002). Antifungal antibiotics. *Appl Microbiol Biotechnol*, 58(1), 46-57. doi:10.1007/s002530100822
- Hibbett, D. S., Binder, M., Bischoff, J. F., Blackwell, M., Cannon, P. F., Eriksson, O. E., . . . Zhang, N. (2007). A higher-level phylogenetic classification of the Fungi. *Mycol Res*, 111(Pt 5), 509-547. doi:10.1016/j.mycres.2007.03.004
- Hillier, L. W., Coulson, A., Murray, J. I., Bao, Z., Sulston, J. E., & Waterston, R. H. (2005). Genomics in *C. elegans*: so many genes, such a little worm. *Genome Res*, 15(12), 1651-1660. doi:10.1101/gr.3729105
- Hoberg, E. P., Monsen, K. J., Kutz, S., & Blouin, M. S. (1999). Structure, biodiversity, and historical biogeography of nematode faunas in holarctic ruminants: morphological and molecular diagnoses for *Teladorsagia boreoarcticus* n. sp. (Nematoda: Ostertagiinae), a dimorphic cryptic species in muskoxen (*Ovibos moschatus*). *J Parasitol*, 85(5), 910-934. doi:10.2307/3285831
- Illy, C., Quraishi, O., Wang, J., Purisima, E., Vernet, T., & Mort, J. S. (1997). Role of the occluding loop in cathepsin B activity. *J Biol Chem*, 272(2), 1197-1202. doi:10.1074/jbc.272.2.1197
- Ingold, C. T., & Hudson, H. J. (1993). *The Biology of Fungi* (6th ed.). Torquay, UK: Springer, Dordrecht.
- Jackson, F., & Coop, R. L. (2000). The development of anthelmintic resistance in sheep nematodes. *Parasitology*, 120 Suppl, S95-107. doi:10.1017/s0031182099005740
- Jalanka, J., Hillamaa, A., Satokari, R., Mattila, E., Anttila, V. J., & Arkkila, P. (2018). The long-term effects of faecal microbiota transplantation for gastrointestinal symptoms and general health in patients with recurrent *Clostridium difficile* infection. *Aliment Pharmacol Ther*, 47(3), 371-379. doi:10.1111/apt.14443
- Janis, C. M. (1993). Tertiary Mammal Evolution in the Context of Changing Climates, Vegetation, and Tectonic Events. *Annu Rev Ecol Evol Syst*, 24(1), 467-500. doi:DOI 10.1146/annurev.es.24.110193.002343
- Johnston, C. J. C., Smyth, D. J., Kodali, R. B., White, M. P. J., Harcus, Y., Filbey, K. J., . . . Maizels, R. M. (2017). A structurally distinct TGF-beta mimic from an intestinal helminth parasite

- potently induces regulatory T cells. *Nat Commun*, 8(1), 1741. doi:10.1038/s41467-017-01886-6
- Juul, S., Izquierdo, F., Hurst, A., Dai, X., Wright, A., Kulesha, E., . . . Turner, D. J. (2015). What's in my pot? Real-time species identification on the MinION™. *bioRxiv*, 030742. doi:10.1101/030742
- Kamra, D. N. (2005). Rumen microbial ecosystem. *Curr Sci*, 89(1), 124-135. Retrieved from <http://www.jstor.org/stable/24110438>
- Kang, J. M., Ju, H. L., Sohn, W. M., & Na, B. K. (2013). Defining the regulatory and inhibitory elements within the prodomain of CsCF-6, a cathepsin F cysteine protease of *Clonorchis sinensis*. *Mol Biochem Parasitol*, 190(2), 92-96. doi:10.1016/j.molbiopara.2013.07.001
- Kang, T. H., Yun, D. H., Lee, E. H., Chung, Y. B., Bae, Y. A., Chung, J. Y., . . . Kong, Y. (2004). A cathepsin F of adult *Clonorchis sinensis* and its phylogenetic conservation in trematodes. *Parasitology*, 128(Pt 2), 195-207. doi:10.1017/s0031182003004335
- Karp, A., Isaac, P. G., & Ingram, D. S. (1998). Isolation of Nucleic Acids Using Silica-Gel Based Membranes: Methods Based on the Use of QIAamp Spin Columns. In A. Karp, P. G. Isaac, & D. S. Ingram (Eds.), *Molecular Tools for Screening Biodiversity* (pp. 59-63). Dordrecht, Netherlands: Springer Netherlands.
- Kasianowicz, J. J., Brandin, E., Branton, D., & Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proc Natl Acad Sci U S A*, 93(24), 13770-13773. doi:10.1073/pnas.93.24.13770
- Kchouk, M. E., Gibrat, J. F., & Elloumi, M. (2017). Generations of Sequencing Technologies: From First to Next Generation. *Biol Med*, 09(03), 1-8. doi:10.4172/0974-8369.1000395
- Keegan, J. D., Good, B., de Waal, T., Fanning, J., & Keane, O. M. (2017). Genetic basis of benzimidazole resistance in *Teladorsagia circumcincta* in Ireland. *Ir Vet J*, 70(1), 8. doi:10.1186/s13620-017-0087-8
- Kemp, B. M., & Smith, D. G. (2005). Use of bleach to eliminate contaminating DNA from the surface of bones and teeth. *Forensic Sci Int*, 154(1), 53-61. doi:10.1016/j.forsciint.2004.11.017
- Kirk, P. M., Cannon, P. F., Minter, D. W., & Stalpers, J. A. (2008). *Dictionary of the Fungi* (10th ed.). Wallingford, UK: CAB International.
- Knox, D. P., Redmond, D. L., Newlands, G. F., Skuce, P. J., Pettit, D., & Smith, W. D. (2003). The nature and prospects for gut membrane proteins as vaccine candidates for *Haemonchus contortus* and other ruminant trichostrongyloids. *Int J Parasitol*, 33(11), 1129-1137. doi:10.1016/S0020-7519(03)00167-X

- Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*, 27(5), 722-736. doi:10.1101/gr.215087.116
- Koski, K. G., & Scott, M. E. (2003). Gastrointestinal nematodes, trace elements, and immunity. *J Trace Elem Exp Med*, 16(4), 237-251. doi:10.1002/jtra.10043
- Kraft, F., & Kurth, I. (2019). Long-read sequencing in human genetics. *Med Genet*, 31(2), 198-204. doi:10.1007/s11825-019-0249-z
- Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*, 227(5259), 680-685. doi:10.1038/227680a0
- Laing, R., Kikuchi, T., Martinelli, A., Tsai, I. J., Beech, R. N., Redman, E., . . . Cotton, J. A. (2013). The genome and transcriptome of *Haemonchus contortus*, a key model parasite for drug and vaccine discovery. *Genome Biol*, 14(8), R88. doi:10.1186/gb-2013-14-8-r88
- Lane, J., Jubb, T., Shephard, R., Webb-Ware, J., Fordyce, G., & GHD Pty Ltd. (2015). *Final report: priority list of endemic diseases for the red meat industries* (B.AHE.0010). Retrieved from <https://era.daf.qld.gov.au/id/eprint/5030/>
- Law, R. H., Smooker, P. M., Irving, J. A., Piedrafita, D., Ponting, R., Kennedy, N. J., . . . Spithill, T. W. (2003). Cloning and expression of the major secreted cathepsin B-like protein from juvenile *Fasciola hepatica* and analysis of immunogenicity following liver fluke infection. *Infect Immun*, 71(12), 6921-6932. doi:10.1128/IAI.71.12.6921-6932.2003
- Leathwick, D. M., & Besier, R. B. (2014). The management of anthelmintic resistance in grazing ruminants in Australasia--strategies and experiences. *Vet Parasitol*, 204(1-2), 44-54. doi:10.1016/j.vetpar.2013.12.022
- Leathwick, D. M., Ganesh, S., & Waghorn, T. S. (2015). Evidence for reversion towards anthelmintic susceptibility in *Teladorsagia circumcincta* in response to resistance management programmes. *Int J Parasitol Drugs Drug Resist*, 5(1), 9-15. doi:10.1016/j.ijpddr.2015.01.001
- Lee, H. J., Jung, J. Y., Oh, Y. K., Lee, S., Madsen, E. L., & Jeon, C. O. (2012). Comparative survey of rumen microbial communities and metabolites across one caprine and three bovine groups, using bar-coded pyrosequencing and ¹H nuclear magnetic resonance spectroscopy. *Appl Environ Microbiol*, 78(17), 5983-5993. doi:10.1128/AEM.00104-12
- Leinonen, R., Sugawara, H., & Shumway, M. (2011). The Sequence Read Archive. *Nucleic Acids Res*, 39(Database issue), D19-D21. doi:10.1093/nar/gkq1019
- Leroy, S., Duperray, C., & Morand, S. (2003). Flow cytometry for parasite nematode genome size measurement. *Mol Biochem Parasitol*, 128(1), 91-93. doi:10.1016/s0166-6851(03)00023-9

- Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., & Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, 299(5607), 682-686. doi:10.1126/science.1079700
- Ley, R. E., Peterson, D. A., & Gordon, J. I. (2006). Ecological and evolutionary forces shaping microbial diversity in the human intestine. *Cell*, 124(4), 837-848. doi:10.1016/j.cell.2006.02.017
- Li, R. W., Li, C., & Gasbarre, L. C. (2011). The vitamin D receptor and inducible nitric oxide synthase associated pathways in acquired resistance to *Cooperia oncophora* infection in cattle. *Vet Res*, 42(1), 48. doi:10.1186/1297-9716-42-48
- Li, Y., Wang, K., Xie, H., Wang, D. W., Xu, C. L., Huang, X., . . . Li, D. L. (2015). Cathepsin B Cysteine Proteinase is Essential for the Development and Pathogenesis of the Plant Parasitic Nematode *Radopholus similis*. *Int J Biol Sci*, 11(9), 1073-1087. doi:10.7150/ijbs.12065
- Li, Y. Y., Fang, J., & Ao, G. Z. (2017). Cathepsin B and L inhibitors: a patent review (2010 - present). *Expert Opin Ther Pat*, 27(6), 643-656. doi:10.1080/13543776.2017.1272572
- Lienhard, A., & Schaffer, S. (2019). Extracting the invisible: obtaining high quality DNA is a challenging task in small arthropods. *PeerJ*, 7, e6753. doi:10.7717/peerj.6753
- Linnevers, C. J., Mcgrath, M. E., Armstrong, A., Mistry, F. R., Barnes, M. G., Klaus, J. L., . . . Brömme, D. (1997). Expression of human cathepsin K in *Pichia pastoris* and preliminary crystallographic studies of an inhibitor complex. *Protein Sci*, 6(4), 919-921. doi:10.1002/pro.5560060421
- Linville, J. G., & Wells, J. D. (2002). Surface sterilization of a maggot using bleach does not interfere with mitochondrial DNA analysis of crop contents. *J Forensic Sci*, 47(5), 1055-1059. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/12353545>
- Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., . . . Law, M. (2012). Comparison of next-generation sequencing systems. *J Biomed Biotechnol*, 2012, 251364. doi:10.1155/2012/251364
- Lopes, L. D., de Souza Lima, A. O., Taketani, R. G., Darias, P., da Silva, L. R., Romagnoli, E. M., . . . Mendes, R. (2015). Exploring the sheep rumen microbiome for carbohydrate-active enzymes. *Anton Leeuw Int J G*, 108(1), 15-30. doi:10.1007/s10482-015-0459-6
- Luque, A., Walker, L. R., Pedley, J. C., Pedley, K. C., Hillrichs, K., Simpson, H. V., & Simcock, D. C. (2010). *Teladorsagia circumcincta*: survival of adults in vitro is enhanced by the presence of a mammalian cell line. *Exp Parasitol*, 124(2), 247-251. doi:10.1016/j.exppara.2009.10.002
- Lutzoni, F., Kauff, F., Cox, C. J., McLaughlin, D., Celio, G., Dentinger, B., . . . Vilgalys, R. (2004). Assembling the fungal tree of life: progress, classification, and evolution of subcellular traits. *Am J Bot*, 91(10), 1446-1480. doi:10.3732/ajb.91.10.1446

- Mahajan, G. B., & Balachandran, L. (2012). Antibacterial agents from actinomycetes - a review. *Front Biosci*, 4, 240-253. doi:10.2741/373
- Maizels, R. M., Smits, H. H., & McSorley, H. J. (2018). Modulation of Host Immunity by Helminths: The Expanding Repertoire of Parasite Effector Molecules. *Immunity*, 49(5), 801-818. doi:10.1016/j.immuni.2018.10.016
- Marchiondo, A. A., Cruthers, L. R., & Reinemeyer, C. R. (2019). Nematoda. In A. A. Marchiondo, L. R. Cruthers, & J. J. Fourie (Eds.), *Parasiticide Screening, Volume 2* (pp. 135-335). London, UK: Academic Press.
- Martin, R. J., Robertson, A. P., & Choudhary, S. (2021). Ivermectin: An Anthelmintic, an Insecticide, and Much More. *Trends Parasitol*, 37(1), 48-64. doi:10.1016/j.pt.2020.10.005
- Martinez-Valladares, M., Donnan, A., Geldhof, P., Jackson, F., Rojo-Vazquez, F. A., & Skuce, P. (2012a). Pyrosequencing analysis of the beta-tubulin gene in Spanish *Teladorsagia circumcincta* field isolates. *Vet Parasitol*, 184(2-4), 371-376. doi:10.1016/j.vetpar.2011.09.009
- Martinez-Valladares, M., Famularo, M. R., Fernandez-Pato, N., Cordero-Perez, C., Castanon-Ordonez, L., & Rojo-Vazquez, F. A. (2012b). Characterization of a multidrug resistant *Teladorsagia circumcincta* isolate from Spain. *Parasitol Res*, 110(5), 2083-2087. doi:10.1007/s00436-011-2753-1
- Martinez-Valladares, M., Valderas-Garcia, E., Gandasegui, J., Skuce, P., Morrison, A., Castilla Gomez de Agüero, V., . . . Rojo-Vazquez, F. A. (2020). *Teladorsagia circumcincta* beta tubulin: the presence of the E198L polymorphism on its own is associated with benzimidazole resistance. *Parasit Vectors*, 13(1), 453. doi:10.1186/s13071-020-04320-x
- Matlock, B. (2015). *Assessment of nucleic acid purity. Technical Note 52646*. Retrieved from <https://assets.thermofisher.com/TFS-Assets/CAD/Product-Bulletins/TN52646-E-0215M-NucleicAcid.pdf>
- Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Proc Natl Acad Sci U S A*, 74(2), 560-564. doi:10.1073/pnas.74.2.560
- Mayson, B. E., Kilburn, D. G., Zamost, B. L., Raymond, C. K., & Lesnicki, G. J. (2003). Effects of methanol concentration on expression levels of recombinant protein in fed-batch cultures of *Pichia methanolica*. *Biotechnol Bioeng*, 81(3), 291-298. doi:10.1002/bit.10464
- McCririe, L., Bairden, K., Britton, C., Buitkamp, J., McKeand, J. B., & Stear, M. J. (1997). Heterogeneity in the recognition of *Ostertagia circumcincta* antigens by serum antibody from mature, infected sheep. *Parasite Immunol*, 19(5), 235-242. doi:10.1046/j.1365-3024.1997.d01-202.x

- McLoughlin, S., Spillane, C., Claffey, N., Smith, P. E., O'Rourke, T., Diskin, M. G., & Waters, S. M. (2020). Rumen Microbiome Composition Is Altered in Sheep Divergent in Feed Efficiency. *Front Microbiol*, 11(1981), 1981. doi:10.3389/fmicb.2020.01981
- Mello, A., Murat, C., & Bonfante, P. (2006). Truffles: much more than a prized and local fungal delicacy. *FEMS Microbiol Lett*, 260(1), 1-8. doi:10.1111/j.1574-6968.2006.00252.x
- Miller, C. M., Waghorn, T. S., Leathwick, D. M., Candy, P. M., Oliver, A. M., & Watson, T. G. (2012). The production cost of anthelmintic resistance in lambs. *Vet Parasitol*, 186(3-4), 376-381. doi:10.1016/j.vetpar.2011.11.063
- Moorthie, S., Mattocks, C. J., & Wright, C. F. (2011). Review of massively parallel DNA sequencing technologies. *Hugo J*, 5(1-4), 1-12. doi:10.1007/s11568-011-9156-3
- Mortimer, R., & Polsinelli, M. (1999). On the origins of wine yeast. *Res Microbiol*, 150(3), 199-204. doi:10.1016/s0923-2508(99)80036-9
- Muhammad, A., Ahmed, H., Iqbal, M. N., & Qayyum, M. (2015). Detection of Multiple Anthelmintic Resistance of *Haemonchus contortus* and *Teladorsagia circumcincta* in Sheep and Goats of Northern Punjab, Pakistan. *Kafkas Univ Vet Fak Derg*, 21(3), 389-395. doi:10.9775/kvfd.2014.12581
- Murphy, L., Eckersall, P. D., Bishop, S. C., Pettit, J. J., Huntley, J. F., Burchmore, R., & Stear, M. J. (2010). Genetic variation among lambs in peripheral IgE activity against the larval stages of *Teladorsagia circumcincta*. *Parasitology*, 137(8), 1249-1260. doi:10.1017/S0031182010000028
- Nari, A., Salles, J., Gil, A., Waller, P. J., & Hansen, J. W. (1996). The prevalence of anthelmintic resistance in nematode parasites of sheep in southern Latin America: Uruguay. *Vet Parasitol*, 62(3-4), 213-222. doi:10.1016/0304-4017(95)00908-6
- Nath, S., Mallick, S. K., & Jha, S. (2014). An improved method of genome size estimation by flow cytometry in five mucilaginous species of Hyacinthaceae. *Cytometry A*, 85(10), 833-840. doi:10.1002/cyto.a.22489
- Nguti, R., Janssen, P., Rowlands, G. J., Audho, J. O., & Baker, R. L. (2003). Survival of Red Maasai, Dorper and crossbred lambs in the sub-humid tropics. *Anim Sci*, 76(1), 3-17. doi:10.1017/S1357729800053261
- Nguyen, V. A., Carey, L. M., Giummarra, L., Faou, P., Cooke, I., Howells, D. W., . . . Crewther, S. G. (2016). A Pathway Proteomic Profile of Ischemic Stroke Survivors Reveals Innate Immune Dysfunction in Association with Mild Symptoms of Depression - A Pilot Study. *Front Neurol*, 7(85), 85. doi:10.3389/fneur.2016.00085
- Nimbkar, C., Ghalsasi, P. M., Swan, A. A., Walkden-Brown, S. W., & Kahn, L. P. (2003). Evaluation of growth rates and resistance to nematodes of Deccani and Bannur lambs and their crosses with Garole. *Anim Sci*, 76(3), 503-515. doi:10.1017/S1357729800058720

- Nisbet, A. J., McNeilly, T. N., Wildblood, L. A., Morrison, A. A., Bartley, D. J., Bartley, Y., . . . Matthews, J. B. (2013). Successful immunization against a parasitic nematode by vaccination with recombinant proteins. *Vaccine*, 31(37), 4017-4023. doi:10.1016/j.vaccine.2013.05.026
- Novinec, M., Pavšič, M., & Lenarčič, B. (2012). A simple and efficient protocol for the production of recombinant cathepsin V and other cysteine cathepsins in soluble form in *Escherichia coli*. *Protein Expr Purif*, 82(1), 1-5. doi:10.1016/j.pep.2011.11.002
- O'Connor, L. J., Walkden-Brown, S. W., & Kahn, L. P. (2006). Ecology of the free-living stages of major trichostrongylid parasites of sheep. *Vet Parasitol*, 142(1-2), 1-15. doi:10.1016/j.vetpar.2006.08.035
- Palevich, N., Maclean, P. H., Mitreva, M., Scott, R., & Leathwick, D. (2019). The complete mitochondrial genome of the New Zealand parasitic roundworm *Teladorsagia circumcincta* (Trichostrongyloidea: Haemonchidae) field strain NZ_Teci_NP. *Mitochondrial DNA B Resour*, 4(2), 2869-2871. doi:10.1080/23802359.2019.1660241
- Pandey, V. S., Chaer, A., & Dakkak, A. (1993). Effect of temperature and relative humidity on survival of eggs and infective larvae of *Ostertagia circumcincta*. *Vet Parasitol*, 49(2-4), 219-227. doi:10.1016/0304-4017(93)90121-3
- Papadopoulos, E., Gallidis, E., & Ptochos, S. (2012). Anthelmintic resistance in sheep in Europe: a selected review. *Vet Parasitol*, 189(1), 85-88. doi:10.1016/j.vetpar.2012.03.036
- Papadopoulos, E., Himonas, C., & Coles, G. C. (2001). Drought and flock isolation may enhance the development of anthelmintic resistance in nematodes. *Vet Parasitol*, 97(4), 253-259. doi:10.1016/S0304-4017(01)00435-6
- Parkinson, J., Mitreva, M., Whitton, C., Thomson, M., Daub, J., Martin, J., . . . Blaxter, M. L. (2004). A transcriptomic analysis of the phylum Nematoda. *Nat Genet*, 36(12), 1259-1267. doi:10.1038/ng1472
- Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9), 1061-1067. doi:10.1093/bioinformatics/btm071
- Pekin, D., Skhiri, Y., Baret, J. C., Le Corre, D., Mazutis, L., Salem, C. B., . . . Taly, V. (2011). Quantitative and sensitive detection of rare mutations using droplet-based microfluidics. *Lab Chip*, 11(13), 2156-2166. doi:10.1039/c1lc20128j
- Perumbakkam, S., Mitchell, E. A., & Morrie Craig, A. (2011). Changes to the rumen bacterial population of sheep with the addition of 2,4,6-trinitrotoluene to their diet. *Anton Leeuw Int J G*, 99(2), 231-240. doi:10.1007/s10482-010-9481-x

- Pillai, S., Gopalan, V., & Lam, A. K. (2017). Review of sequencing platforms and their applications in pheochromocytoma and paragangliomas. *Crit Rev Oncol Hematol*, 116, 58-67. doi:10.1016/j.critrevonc.2017.05.005
- Pitta, D. W., Pinchak, W. E., Dowd, S. E., Osterstock, J., Gontcharova, V., Youn, E., . . . Malinowski, D. P. (2010). Rumen Bacterial Diversity Dynamics Associated with Changing from Bermudagrass Hay to Grazed Winter Wheat Diets. *Microb Ecol*, 59(3), 511-522. doi:10.1007/s00248-009-9609-6
- Poinar, G. (2012). Nematoda (Roundworms). In *eLS*. Chichester, UK: John Wiley & Sons.
- Puzer, L., Cotrin, S. S., Alves, M. F. M., Egborge, T., Araújo, M. S., Juliano, M. A., . . . Carmona, A. K. (2004). Comparative substrate specificity analysis of recombinant human cathepsin V and cathepsin L. *Arch Biochem Biophys*, 430(2), 274-283. doi:10.1016/j.abb.2004.07.006
- Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., . . . Gu, Y. (2012). A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, 13(1), 341. doi:10.1186/1471-2164-13-341
- Rajilic-Stojanovic, M., & de Vos, W. M. (2014). The first 1000 cultured species of the human gastrointestinal microbiota. *FEMS Microbiol Rev*, 38(5), 996-1047. doi:10.1111/1574-6976.12075
- Rappsilber, J., Mann, M., & Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat Protoc*, 2(8), 1896-1906. doi:10.1038/nprot.2007.261
- Redmond, D. L., Smith, S. K., Halliday, A., Smith, W. D., Jackson, F., Knox, D. P., & Matthews, J. B. (2006). An immunogenic cathepsin F secreted by the parasitic stages of *Teladorsagia circumcincta*. *Int J Parasitol*, 36(3), 277-286. doi:10.1016/j.ijpara.2005.10.011
- Renoux, G. (1980). The general immunopharmacology of levamisole. *Drugs*, 20(2), 89-99. doi:10.2165/00003495-198020020-00001
- Rivera-Mariani, F. E., & Bolaños-Rosero, B. (2011). Allergenicity of airborne basidiospores and ascospores: need for further studies. *Aerobiologia*, 28(2), 83-97. doi:10.1007/s10453-011-9234-y
- Riviere, J. E., & Papich, M. G. (2009). *Veterinary Pharmacology and Therapeutics* (J. E. Riviere & M. G. Papich Eds. Illustrated ed.). Hoboken, USA: John Wiley & Sons.
- Roger, A. J., Munoz-Gomez, S. A., & Kamikawa, R. (2017). The Origin and Diversification of Mitochondria. *Curr Biol*, 27(21), R1177-R1192. doi:10.1016/j.cub.2017.09.015
- Rousk, J., Baath, E., Brookes, P. C., Lauber, C. L., Lozupone, C., Caporaso, J. G., . . . Fierer, N. (2010). Soil bacterial and fungal communities across a pH gradient in an arable soil. *ISME J*, 4(10), 1340-1351. doi:10.1038/ismej.2010.58

- Sanchez-Ayala, J. R., Cruz-Mendoza, I., Figueroa-Castillo, J. A., & Vital-Garcia, C. (2018). First report of *Libyostrongylus douglassii* (Strongylida: Trichostrongylidae) in ostriches (*Struthio camelus*) from Mexico. *Vet Parasitol: Reg Stud Rep*, 12, 31-34. doi:10.1016/j.vprsr.2018.01.007
- Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74(12), 5463-5467. doi:10.1073/pnas.74.12.5463
- Savage, R. J. G., & Long, M. R. (1986). *Mammal evolution : an illustrated guide*. New York, USA: Facts on File Publications.
- Schoch, C. L., Sung, G. H., Lopez-Giraldez, F., Townsend, J. P., Miadlikowska, J., Hofstetter, V., . . . Spatafora, J. W. (2009). The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Syst Biol*, 58(2), 224-239. doi:10.1093/sysbio/syp020
- Schwaiger, F. W., Gostomski, D., Stear, M. J., Duncan, J. L., McKellar, Q. A., Epplen, J. T., & Buitkamp, J. (1995). An ovine Major histocompatibility complex DRB1 allele is associated with low faecal egg counts following natural, predominantly *Ostertagia circumcincta* infection. *Int J Parasitol*, 25(7), 815-822. doi:10.1016/0020-7519(94)00216-b
- Schwarz, E. M., Korhonen, P. K., Campbell, B. E., Young, N. D., Jex, A. R., Jabbar, A., . . . Gasser, R. B. (2013). The genome and developmental transcriptome of the strongylid nematode *Haemonchus contortus*. *Genome Biol*, 14(8), R89. doi:10.1186/gb-2013-14-8-r89. (Accession No. 23985341)
- Seesao, Y., Audebert, C., Verrez-Bagnis, V., Merlin, S., Jerome, M., Viscogliosi, E., . . . Gay, M. (2014). Monitoring of four DNA extraction methods upstream of high-throughput sequencing of Anisakidae nematodes. *J Microbiol Methods*, 102, 69-72. doi:10.1016/j.mimet.2014.05.004
- Setsuda, A., Kato, E., Sakaguchi, S., Tamemasa, S., Ozawa, S., & Sato, H. (2019). *Chabaudstrongylus ninhiae* (Trichostrongylidae: Cooperiinae) and *Oesophagostomum muntiacum* (Chabertiidae: Oesophagostominae) in feral alien Reeves's muntjacs on Izu-Oshima Island, Tokyo, Japan. *J Helminthol*, 94, e48. doi:10.1017/S0022149X19000245
- Silva, N., Igrejas, G., Goncalves, A., & Poeta, P. (2012). Commensal gut bacteria: distribution of Enterococcus species and prevalence of *Escherichia coli* phylogenetic groups in animals and humans in Portugal. *Ann Microbiol*, 62(2), 449-459. doi:10.1007/s13213-011-0308-4
- Simbolo, M., Gottardi, M., Corbo, V., Fassan, M., Mafficini, A., Malpeli, G., . . . Scarpa, A. (2013). DNA qualification workflow for next generation sequencing of histopathological samples. *PLoS One*, 8(6), e62692. doi:10.1371/journal.pone.0062692

- Sloan, S., Jenvey, C., Cairns, C., & Stear, M. (2020). Cathepsin F of *Teladorsagia circumcincta* is a recently evolved cysteine protease. *Evol Bioinform Online*, 16, 1176934320962521. doi:10.1177/1176934320962521
- Sloan, S., Jenvey, C. J., Piedrafita, D., Preston, S., & Stear, M. J. (2021). Comparative evaluation of different molecular methods for DNA extraction from individual *Teladorsagia circumcincta* nematodes. *BMC Biotechnol*, 21(1), 35. doi:10.1186/s12896-021-00695-6
- Smith, A. M., Dowd, A. J., Heffernan, M., Robertson, C. D., & Dalton, J. P. (1993). *Fasciola hepatica*: a secreted cathepsin L-like proteinase cleaves host immunoglobulin. *Int J Parasitol*, 23(8), 977-983. doi:10.1016/0020-7519(93)90117-h
- Smith, S. M., & Gottesman, M. M. (1989). Activity and Deletion Analysis of Recombinant Human Cathepsin L Expressed in *Escherichia coli**. *J Biol Chem*, 264(34), 20487-20495. doi:10.1016/S0021-9258(19)47088-9
- Smith, W. D., Jackson, F., Jackson, E., Graham, R., Williams, J., Willadsen, S. M., & Fehilly, C. B. (1986). Transfer of immunity to *Ostertagia circumcincta* and IgA memory between identical sheep by lymphocytes collected from gastric lymph. *Res Vet Sci*, 41(3), 300-306. doi:10.1016/S0034-5288(18)30620-9
- Smith, W. D., Jackson, F., Jackson, E., & Williams, J. (1983). Local immunity and *Ostertagia circumcincta*: changes in the gastric lymph of immune sheep after a challenge infection. *J Comp Pathol*, 93(3), 479-488. doi:10.1016/0021-9975(83)90035-x
- Smith, W. D., Jackson, F., Jackson, E., & Williams, J. (1985). Age immunity to *Ostertagia circumcincta*: comparison of the local immune responses of 4 1/2- and 10-month-old lambs. *J Comp Pathol*, 95(2), 235-245. doi:10.1016/0021-9975(85)90010-6
- Somoza, J. R., Palmer, J. T., & Ho, J. D. (2002). The crystal structure of human cathepsin F and its implications for the development of novel immunomodulators. *J Mol Biol*, 322(3), 559-568. doi:10.1016/S0022-2836(02)00780-5
- Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19(suppl_2), ii215-ii225. doi:10.1093/bioinformatics/btg1080
- Stear, M. J., Bairden, K., Bishop, S. C., Buitkamp, J., Duncan, J. L., Gettinby, G., . . . Murray, M. (1997). The genetic basis of resistance to *Ostertagia circumcincta* in lambs. *Vet J*, 154(2), 111-119. doi:10.1016/s1090-0233(97)80049-4
- Stear, M. J., Bairden, K., Bishop, S. C., Gettinby, G., McKellar, Q. A., Park, M., . . . Wallace, D. S. (1998). The processes influencing the distribution of parasitic nematodes among naturally infected lambs. *Parasitology*, 117 (Pt 2)(2), 165-171. doi:10.1017/s0031182098002868
- Stear, M. J., Bairden, K., Duncan, J. L., Eckersall, P. D., Fishwick, G., Graham, P. A., . . . Wallace, D. S. (2000). The influence of relative resistance and urea-supplementation on deliberate

- infection with *Teladorsagia circumcincta* during winter. *Vet Parasitol*, 94(1-2), 45-54.
doi:10.1016/s0304-4017(00)00370-8
- Stear, M. J., & Bishop, S. C. (1999a). The curvilinear relationship between worm length and fecundity of *Teladorsagia circumcincta*. *Int J Parasitol*, 29(5), 777-780.
doi:10.1016/s0020-7519(99)00019-3
- Stear, M. J., Bishop, S. C., Doligalska, M., Duncan, J. L., Holmes, P. H., Irvine, J., . . . Murray, M. (1995a). Regulation of egg production, worm burden, worm length and worm fecundity by host responses in sheep infected with *Ostertagia circumcincta*. *Parasite Immunol*, 17(12), 643-652. doi:10.1111/j.1365-3024.1995.tb01010.x
- Stear, M. J., Bishop, S. C., Duncan, J. L., McKellar, Q. A., & Murray, M. (1995b). The repeatability of faecal egg counts, peripheral eosinophil counts, and plasma pepsinogen concentrations during deliberate infections with *Ostertagia circumcincta*. *Int J Parasitol*, 25(3), 375-380. doi:10.1016/0020-7519(94)00136-c
- Stear, M. J., Bishop, S. C., Henderson, N. G., & Scott, I. (2003). A key mechanism of pathogenesis in sheep infected with the nematode *Teladorsagia circumcincta*. *Anim Health Res Rev*, 4(1), 45-52. doi:10.1079/ahrr200351
- Stear, M. J., Doligalska, M., & Donskow-Schmelter, K. (2007). Alternatives to anthelmintics for the control of nematodes in livestock. *Parasitology*, 134(Pt 2), 139-151.
doi:10.1017/S0031182006001557
- Stear, M. J., Singleton, D., & Matthews, L. (2011). An evolutionary perspective on gastrointestinal nematodes of sheep. *J Helminthol*, 85(2), 113-120. doi:10.1017/S0022149X11000058
- Stear, M. J., Strain, S., & Bishop, S. C. (1999b). Mechanisms underlying resistance to nematode infection. *Int J Parasitol*, 29(1), 51-56; discussion 73-55. doi:10.1016/s0020-7519(98)00179-9
- Stebbins, G. L. (1981). Coevolution of Grasses and Herbivores. *Ann Missouri Bot Gard*, 68(1), 75-86. doi:Doi 10.2307/2398811
- Stevenson, L. A., Gasser, R. B., & Chilton, N. B. (1996). The ITS-2 rDNA of *Teladorsagia circumcincta*, *T-trifurcata* and *T-davtiani* (Nematoda: Trichostrongylidae) indicates that these taxa are one species. *In J Parasitol*, 26(10), 1123-1126. doi:Doi 10.1016/S0020-7519(96)00064-1
- Su, Z., Ning, B., Fang, H., Hong, H., Perkins, R., Tong, W., & Shi, L. (2011). Next-generation sequencing and its applications in molecular diagnostics. *Expert Rev Mol Diagn*, 11(3), 333-343. doi:10.1586/erm.11.3
- Tang, Y. T., Gao, X., Rosa, B. A., Abubucker, S., Hallsworth-Pepin, K., Martin, J., . . . Mitreva, M. (2014). Genome of the human hookworm *Necator americanus*. *Nat Genet*, 46(3), 261-269. doi:10.1038/ng.2875

- Taylor, M. A., Coop, R. L., & Wall, R. L. (2007). *Veterinary parasitology* (3rd ed.). Oxford, UK: Blackwell.
- Thomas, F., Hehemann, J. H., Rebuffet, E., Czejek, M., & Michel, G. (2011). Environmental and gut bacteroidetes: the food connection. *Front Microbiol*, 2(93), 93. doi:10.3389/fmicb.2011.00093
- Toguebaye, B. S., Quilichini, Y., Diagne, P. M., & Marchand, B. (2014). Ultrastructure and development of *Nosema podocotyloidis* n. sp. (Microsporidia), a hyperparasite of *Podocotyloides magnatestis* (Trematoda), a parasite of *Parapristipoma octolineatum* (Teleostei). *Parasite*, 21, 44. doi:10.1051/parasite/2014044
- Towbin, H., Staehelin, T., & Gordon, J. (1979). Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications. *Proc Natl Acad Sci U S A*, 76(9), 4350-4354. doi:10.1073/pnas.76.9.4350
- Traversa, D., Paoletti, B., Otranto, D., & Miller, J. (2007). First report of multiple drug resistance in trichostrongyles affecting sheep under field conditions in Italy. *Parasitol Res*, 101(6), 1713-1716. doi:10.1007/s00436-007-0707-4
- Turk, V., Stoka, V., Vasiljeva, O., Renko, M., Sun, T., Turk, B., & Turk, D. (2012). Cysteine cathepsins: from structure, function and regulation to new frontiers. *Biochim Biophys Acta*, 1824(1), 68-88. doi:10.1016/j.bbapap.2011.10.002
- Urquhart, G. M., Jarrett, W. F., Jennings, F. W., McIntyre, W. I., & Mulligan, W. (1966). Immunity to *Haemonchus contortus* infection: relationship between age and successful vaccination with irradiated larvae. *Am J Vet Res*, 27(121), 1645-1648. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/6008311>
- Ventura, M., Canchaya, C., Tauch, A., Chandra, G., Fitzgerald, G. F., Chater, K. F., & van Sinderen, D. (2007). Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiol Mol Biol Rev*, 71(3), 495-548. doi:10.1128/MMBR.00005-07
- Venturina, V. M., Gossner, A. G., & Hopkins, J. (2013). The immunology and genetics of resistance of sheep to *Teladorsagia circumcincta*. *Vet Res Commun*, 37(2), 171-181. doi:10.1007/s11259-013-9559-9
- Vidak, E., Javorsek, U., Vizovisek, M., & Turk, B. (2019). Cysteine Cathepsins and their Extracellular Roles: Shaping the Microenvironment. *Cells*, 8(3). doi:10.3390/cells8030264
- Vilgis, S., & Deigner, H. (2018). Sequencing in Precision Medicine. In H. Deigner & M. Kohl (Eds.), *Precision Medicine* (pp. 79-101). Cambridge, USA: Academic Press.
- Waller, P. J., Schwan, O., Ljungstrom, B. L., Rydzik, A., & Yeates, G. W. (2004a). Evaluation of biological control of sheep parasites using *Duddingtonia flagrans* under commercial farming conditions on the island of Gotland, Sweden. *Vet Parasitol*, 126(3), 299-315. doi:10.1016/j.vetpar.2004.08.008

- Waller, P. J., & Thamsborg, S. M. (2004b). Nematode control in 'green' ruminant production systems. *Trends Parasitol*, 20(10), 493-497. doi:10.1016/j.pt.2004.07.012
- Walsh, P. S., Metzger, D. A., & Higuchi, R. (1991). Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *Biotechniques*, 10(4), 506-513. doi:10.2144/000114018
- Wang, B., Shi, G. P., Yao, P. M., Li, Z., Chapman, H. A., & Bromme, D. (1998). Human cathepsin F. Molecular cloning, functional expression, tissue localization, and enzymatic characterization. *J Biol Chem*, 273(48), 32000-32008. doi:10.1074/jbc.273.48.32000
- Wang, J., Liu, J., & Kang, M. (2015). Quantitative testing of the methodology for genome size estimation in plants using flow cytometry: a case study of the *Primulina* genus. *Front Plant Sci*, 6, 354. doi:10.3389/fpls.2015.00354
- Wex, T., Wex, H., & Bromme, D. (1999). The human cathepsin F gene--a fusion product between an ancestral cathepsin and cystatin gene. *Biol Chem*, 380(12), 1439-1442. doi:10.1515/BC.1999.185
- Williams, K. P., Sobral, B. W., & Dickerman, A. W. (2007). A robust species tree for the alphaproteobacteria. *J Bacteriol*, 189(13), 4578-4586. doi:10.1128/JB.00269-07
- Wood, W. (2002a). *Caenorhabditis*. In *Encyclopedia of Molecular Biology*. Hoboken, USA: John Wiley & Sons.
- Wood, W. (2002b). *Nematodes*. In *Encyclopedia of Molecular Biology*. Hoboken, USA: John Wiley & Sons.
- Woolastont, R. R., & Windon, R. G. (2001). Selection of sheep for response to *Trichostrongylus colubriformis* larvae: genetic parameters. *Anim Sci*, 73(1), 41-48. doi:10.1017/s1357729800058033
- Yang, Y., Yang, E., An, Z., & Liu, X. (2007). Evolution of nematode-trapping cells of predatory fungi of the Orbiliaceae based on evidence from rRNA-encoding DNA and multiprotein sequences. *Proc Natl Acad Sci U S A*, 104(20), 8379-8384. doi:10.1073/pnas.0702770104
- Yazwinski, T. A., Goode, L., Moncol, D. J., Morgan, G. W., & Linnerud, A. C. (1979). Parasite resistance in straightbred and crossbred Barbados Blackbelly sheep. *J Anim Sci*, 49(4), 919-926. doi:10.2527/jas1979.494919x

Appendix

S2.1 Chapter 2 Supplementary File 1

Alignment of the mRNA sequence for *Teladorsagia circumcincta* cathepsin F (GenBank accession no. DQ133568) against the *T. circumcincta* genome in WormBase ParaSite (BioProject: PRJNA72569, Taxonomy ID: 45464)

TELCIR_06733

>scaffold:T_circumcincta.14.0.ec.cg.pg:TELCIRDFT_Contig989:6912:20594:-1

GenBank accession no. DQ133568 match highlighted in grey.

TELCIR_06733 predicted exon 1:

19993

ATGATCCGAAAATCTTCACAGATTACTTACCGTACACATGGAAACAATCACATCATTCTGAACCGAATCGT
GAACCTAGTCGCCGAAGGAGTGGATCCAAAGAAGCCATTGCCAGAATCATTGATTGGAGAAAACATGG
CGCAGTGACCGAAGTTAAAGATCAAG 19829

TELCIR_06733 predicted exon 2:

18402

GTCAGTGTGGGAGCTGTTGGGCGTTCTCTACCACAGGAAATATCGAAGGCCAGTGGTTCCTGGCCAGAA
AGGAACTGGTGTGCTTTTCGGAACAGGAACTCCTTGATTGTGATGAGGTTGATTGGGGATGTAATGGTG
GGCTGCCTATCGACGCTTACCA 18243

TELCIR_06733 predicted exon 3:

15436

GGAGATCATGCGGATAGGCGGCTTGGAATCAGAAGACGATTATCCGTACGAGGCAAAAGAAGAGAAAT
GTCACCTTGTCCTGCTCG 15352

TELCIR_06733 predicted exon 4:

14655

ATCAAGCATTGCCAACAGTGTTTCCATAGCATCTTGTTTGGGGGTGTCGTCATGATAGCGAAACCAGAGT
TGATTGGCATGTGCGATGCGAGTTTCGATTCCAACAAAAGTTTTGTAGAGTGCTCTTCGGATTCTGAAGCA
CTTACCATTGCTCTACCTGGTATGGTTTCTACCGCCTGTCGTAGTAGCACTAAAACCTCTGCTGAAGG
AGGCTAGGCTTCTTGGTATCATGCTGCTTGTGAATGGTGATTTCTTCACTGTTTTGTGTCTTGAAGCCCGTC
GTAAAGACTTCTGG 14399

TELCIR_06733 predicted exon 5:

12879

AATATCTCCGTTTACATCAACGGCTCAGTCGAGCTGCCACATGATGAGGAAAGCATGAGAGCGTGGCTAG
TGAATAAGGGACCGATATCGATAG 12786

TELCIR_06733 predicted exon 6:

12563

GAAAAAGAAAAAGAAGAAAGAAATTCAGAAAAGTGAATCTTGGAATTGTCCACGACTACTCATCGC
CTCATCATTTAGGTATCAACGCAGATAATATGATATTCTATAAAAGTGGCATTGCTCGTCCGCGATTCTGT
GATCCAGACGAGCTAAATCACGGTGTTCTATTAGTCGGATACGGTATCGAAGGGAAGAAGCCCTATTGG
ATAATAAAGAACTCTGGGGGTCTGATTGGGGAGAGGGAGGATATTACAG 12305

TELCIR_06733 predicted exon 7:

GATCATTCGTGGGAAGAACGCCTGCGGCCTGAACCAAATGCCAACATCGGCTGTTGTCCAGTAG

TELCIR_06734

>scaffold:T_circumcincta.14.0.ec.cg.pg:TELCIRDFT_Contig989:21272:26430:-1

GenBank accession no. DQ133568 match highlighted in grey.

TELCIR_06734 predicted exon 1:

25830

ATGGCTTACAATAAATATTATACTTTAGGGATGTATGTCTGGTTCCTGTTCTCATCCCGCACTTATTTGCC
GCTGCTGTAAAGCAGGAAGACTCGGGAGAAGTCAAACCATTTGGAAGATGTCCACACGGATTTAGTTGAC
GAGATAAC 25682

TELCIR_06734 predicted exon 2:

14403

CAAAGGCTCTGTCGAATACAGCAGACTCGGTCGATACATCAATCCAAATGAATTGAATGCTTGAATCAG
TTCACCAACTTCATTGAAAG 24314

TELCIR_06734 predicted exon 3:

23930

GCACGGCAAGAGCTACAGCAGCGAAAGTGAAGCTCTAGAACGATTGCAATTTCAAAAGGAATTTGGA
G 23861

TELCIR_06734 predicted exon 4:

22003

GTGATTCGCACTATGCAAGAAAACGAACAGGGAAGTCTGTTTATGGAATCACGCGGTTTCGCTGATCTTT
CACCGGAGGAATTCAAAAAGTATTTTTTTTTATGGCGTGCACTTTCATGAGGATTTCTAA 21871

TELCIR_20397

>scaffold:T_circumcincta.14.0.ec.cg.pg:TELCIRDFT_Contig17435:3209:5551:1

GenBank accession no. DQ133568 match highlighted in grey.

TELCIR_20397 predicted exon 1:

3809 ATG 3811

TELCIR_20397 predicted exon 2:

3970

GATATCGCCGTTTATATCAACGGCTCAGTCGAGCTACCACATGATGAGGAAAAAATGAGGGCATGGCTA
GTGAAGAAGGGGCCGATATCCATAG 4063

TELCIR_20397 predicted exon 3:

4770

GTATCACCGTAGATGACATACAGTTCCATAAAGGCGGCGTTTCTCGTCCGACTACCTGTAGACCATCTTCT
ATGATTCATGGCGCTCTTCTGGTCGGATACGGTGTCGAGAAGAATATACCGTACTGGATTATAAAGAATT
CATGGGGCCCCAATTGGGGAGAGGATGGATATTACAGGTAA 4951

TELCIR_14223

>scaffold:T_circumcincta.14.0.ec.cg.pg:TELCIRDFT_Contig5194:6821:27433:-1

GenBank accession no. DQ133568 match highlighted in grey.

TELCIR_14223 predicted exon 1:

26833

ATGTCTTGTGGACTGCCCCTGTCCAAGGTCGGGAAAAACGCTGCGAGGTACGAACACGCGCTGAGGAG
ATCACAGAACACCGGATAAAGGAGGCAGTGAATAGTTATTCCAACGCAAGCCACAGCGAGGAGCCCTAC
ACGAAGAAGGACCTTCCTGAAAGCGGGGCCAATTAGAG 26657

TELCIR_14223 predicted exon 2:

26537

ATAATTCGCACTGCGCAGGAAAACGATAAGGGAACAGCTATTTACGGAATCAACCAGTTTGCTGATCTTT
CACCGGAGGAATTCAAAAAG 26348

TELCIR_14223 predicted exon 3:

24061

ACTCACCTGCCGCACACATGGAAACAGCCTGATCATCCAAACCGAATCGTGGACTTAGCCGCAGAAGGG
GTGGATCCGAAGGAGCCACTGCCGGAATCGTTCGATTGGAGAGAACATGGTGCAGTGACAAAAGTGAA
AACTGAAG 23917

TELCIR_14223 predicted exon 4:

23578

GTCAGTGTGCAGCCTGCTGGGCATTTTCTGTCACAGGAAATATTGAAGGCCAGTGGTTCCTTGCCAAAAA
GAAGCTTGTATCGCTTTCGGCACAACAGCTCCTTGATTGTGATGTTGTTGATGAGGGATGTAACGGTGGA
TTTCCTCTTGACGCTTACAA 23419

TELCIR_14223 predicted exon 5:

22979

AGAAATCGTTCGAATGGGCGGCTTGGAATCAGAAGACAAGTATCCCTACGAAGCCAAGGCAGAGCAGTG
TCGCCTTGTCCTCATCG 22895

TELCIR_14223 intron between predicted exons 5 and 6:

21023

ATAAGTTTCAGGATATCGCTGTTTATATCAACGGCTCAGTCGAGCTACCACATGATGAAGAAAAAATGAG
GGCANNNNNNNNNNNNNNNNNN 20934

TELCIR_14223 predicted exon 6:

18316

GATGGTGCGTGGGGAGAACGCTTGTCGCATAAACAGATCCCCACGTCAGCTGTTGTCCTATAATCGTTG
CTCAGTCCAGCATCACCGATCGTCGCTCAATTCAACAGCACTATCTTCAACAACATTGTA 18187

TELCIR_14223 predicted exon 7:

9542

GAATGTTTGTCTGGTTCTTGTCTCATCTCGCCGTCGTTTGTAGCTCCGTAAAGCAGAAACACTCTGGA
GAAGCGAAACCGTTGACAGATCCTCATACGGATTTGATTGATGACATAAC 9422

TELCIR_14223 intron between predicted exons 7 and 8:

7823

CGGAGGCTCCATCGAATACACCAGGCTCAGTCGATACATCAGTCCAAATGACTTCGGTGCTTGGAACAA
TTCACCAGCTTCATTGAAAGGTTTCATTCAACGTAAATAACTTCTTTCAA 7704

TELCIR_14223 predicted exon 8:

7580

GCACGGCAAAGTCTACAGGAATGAAAGCGAAGCCCTAAATCGATTCCGGGTCTTCAAAAGAAATCTGGA

GGTACATAACGAAAATTTTGCCCAAAAAATCATTCTCTACTTCTAATCACCTCATTTTCCCGGCACCTTCA
ATTCTCTGGGTTCGTTTGA 7421

TELCIR_19209

>scaffold:T_circumcincta.14.0.ec.cg.pg:TELCIRDFT_Contig13228:7818:10629:-1

GenBank accession no. DQ133568 match highlighted in grey.

Upstream of TELCIR_19209:

10629

TTCTGAATTCCAGATAATTCGCTCTGCGCAGGAAAACGATAAGGGAACAGCTATTTACGGAATCAATCAG
TTTGCTGATCTTTCACCGGAGGAATTTAAAAAGGTATTTCAACAGATGTG 10509

TELCIR_19209 predicted exon 1:

9398

ATGACTTGTCAATTTACGCAGACTCACCTGCCGCGCACATGGAAACAACCTGAGCATCCAAACCGAATCG
TAGACTTAGCCGCAGAAGGGGTGGATCCGAAGGAGCCACTGCCGGAATCGTTCGATTGGAGAGAACAT
GGTGCACTGACAAAAGTGAAAAGTGAAG 9233

TELCIR_19209 predicted exon 2:

8892

GTCAGTGTGCAGCTTGCTGGGCATTTTCTGTACAGGAAATATTGAAGGCCAGTGGTTCCTTGCCAAAAA
GAAACTTGTATCGCTTTCGGCACAACTCCTTGATTGTGATGTTGTTGATGAGGGATGTAACGGTGGA
TTTCCTCTTGACGCTTACAA 8733

TELCIR_19202 predicted exon 3:

8460 AGAAATCGTTCGAATGGGCGGCTTGAATCAGAAGACAAGTAT 8418

Downstream of TELCIR_19209:

7968

NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNCAAGGCAGAGCAGTGTGCGCTTGTC CATCGGTT
AGTATTGGTTTTTTGAAGAAATACAGGTTGTCTCGAAATCATTTGCGCCTACTTG 7849

S2.2 Chapter 2 Supplementary File 2




Alignment of TELCIR_06733 and _06734 translated exons from draft *Teladorsagia circumcincta* genome in WormBase Parasite (BioProject: PRJNA72569) against *T. circumcincta* secreted cathepsin F (Tci-CF-1, GenBank accession no. ABA01328). Conserved residues indicated by a dot; gap indicated by a dash; stop indicated by an asterisk.

				20				40				60	
Tci-CF-1	MS	-----	--LLFLL	IP	HLFAATVKQQ	YSGGVKPLTE	LRTDLIDKKT	KGS	IEFARLG			50	
TELCIR_06734 exon 1	.AYNKYYTLG	MYVW.	.F...	A...E	D..E....	ED	VH...	V.EI-	-----		49	
TELCIR_06734 exon 2	-----	-----	-----	-----	-----	-----	-----	-----	-----	...V.YS	...	10	
TELCIR_06734 intron	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06734 exon 4	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 exon 1	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 exon 2	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 exon 3	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 exon 5	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 exon 6	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 intron	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
				80				100				120	
Tci-CF-1	QHISPKDFGA	WNHFTS	FIER	HDKVYRNES	E	ALKRFGIFKR	NLEIIRSAQE	NDKGTAI	YGI			110	
TELCIR_06734 exon 1	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	49	
TELCIR_06734 exon 2	RY.N.NELN.	.Q..N...	-----	-----	-----	-----	-----	-----	-----	-----	-----	29	
TELCIR_06734 intron	-----	-----	-----	.G.S.SSE..A...	-----	-----	-----	-----	-----	23	
TELCIR_06734 exon 4	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	17	
TELCIR_06733 exon 1	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 exon 2	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 exon 3	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 exon 5	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 exon 6	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 intron	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
				140				160				180	
Tci-CF-1	NQFADLSPEE	FKKTHLPHTW	KQPDHPNRIV	DLAAEGVDPK	EPLPES	FDWR	EHGAVTKVKT					170	
TELCIR_06734 exon 1	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	49	
TELCIR_06734 exon 2	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	29	
TELCIR_06734 intron	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	23	
TELCIR_06734 exon 4	TR.....	...VFF	-----	-----	-----	-----	-----	-----	-----	-----	-----	33	
TELCIR_06733 exon 1	-----	-----	...Y.	...SH.S...	N.V.....	K.....	K.....	E..D	45				
TELCIR_06733 exon 2	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 exon 3	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 exon 5	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 exon 6	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 intron	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
				200				220				240	
Tci-CF-1	EGHCAACWAF	SVTGNIEGQW	FLAKKKLVSL	SAQQLLDCDV	VDEGCNGGFP	LDAYKE	IVRM					230	
TELCIR_06734 exon 1	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	49	
TELCIR_06734 exon 2	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	29	
TELCIR_06734 intron	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	23	
TELCIR_06734 exon 4	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	33	
TELCIR_06733 exon 1	Q.....	...VFF	-----	-----	-----	-----	-----	-----	-----	-----	-----	46	
TELCIR_06733 exon 2	...GS...	.T.....	...R.E...	.E.E....	E..W....	L..I....	..M..I	5					
TELCIR_06733 exon 3	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 exon 5	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 exon 6	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
TELCIR_06733 intron	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
				260				280				300	
Tci-CF-1	GGLEPEDKYP	YEAKAEQCRL	VPSDIAVYIN	GSVELPHDEE	KMRAWLVKKG	PIS	IGITVDD					290	
TELCIR_06734 exon 1	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	49	
TELCIR_06734 exon 2	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	29	
TELCIR_06734 intron	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	23	
TELCIR_06734 exon 4	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	33	
TELCIR_06733 exon 1	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	46	
TELCIR_06733 exon 2	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	52	
TELCIR_06733 exon 3	...S..D..	...E.K.H.	.R.....	-----	-----	-----	-----	-----	-----	-----	-----	28	
TELCIR_06733 exon 5	-----	-----	-----	...N.S...	-----	S.....	N...	-----	-----	-----	-----	31	
TELCIR_06733 exon 6	-----	-----	-----	-----	-----	-----	-----	S	..HHL..	NA..N	11		
TELCIR_06733 intron	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
				320				340				360	
Tci-CF-1	IQFYKGGVSR	PTTCRLSSMI	HGALLVGYG	EKNIPYWI	IK	NSWGP	NWGED	GYRMR	VGEN			350	
TELCIR_06734 exon 1	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	49	
TELCIR_06734 exon 2	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	29	
TELCIR_06734 intron	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	23	
TELCIR_06734 exon 4	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	33	
TELCIR_06733 exon 1	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	46	
TELCIR_06733 exon 2	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	52	
TELCIR_06733 exon 3	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	28	
TELCIR_06733 exon 5	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	31	
TELCIR_06733 exon 6	M1...S..IA.	.RF..DPDELN	..V.....	I..GKK...SD...	G.....I..	K..6					
TELCIR_06733 intron	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	-----	
				320				340				360	
Tci-CF-1	ACRINRFPTS	AVVL										364	
TELCIR_06734 exon 1	-----	----										49	
TELCIR_06734 exon 2	-----	----										29	
TELCIR_06734 intron	-----	----										23	
TELCIR_06734 exon 4	-----	----										33	
TELCIR_06733 exon 1	-----	----										46	
TELCIR_06733 exon 2	-----	----										52	
TELCIR_06733 exon 3	-----	----										28	
TELCIR_06733 exon 5	-----	----										31	
TELCIR_06733 exon 6	-----	----										64	
TELCIR_06733 intron	..GL..QM...	...Q	20										

S2.3 Chapter 2 Supplementary File 3

Phyre² secondary structure and disorder prediction, detailed template information, and domain analysis for *T. circumcincta* cathepsin F (GenBank accession no. DQ133568).


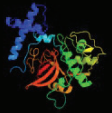

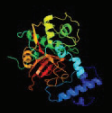

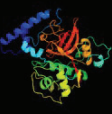

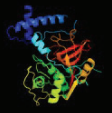

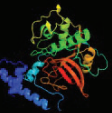

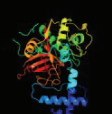

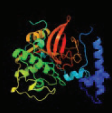



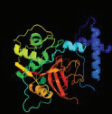

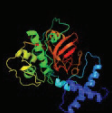

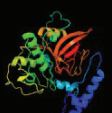



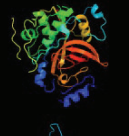
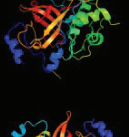
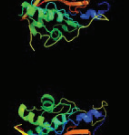
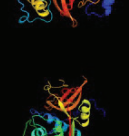

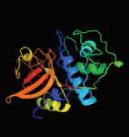
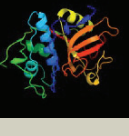
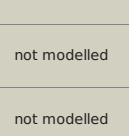
Confidence Key
High(9)  Low (0)
? Disordered (16%)
 Alpha helix (37%)
 Beta strand (15%)

Phyre2

Detailed template information

Unique Job ID 93655db4c2450b0b

#	Template	Alignment Coverage	3D Model	Confidence	% i.d.	Template Information
1	d7pcka	 Alignment		100.0	32	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
2	c5egwA	 Alignment		100.0	33	PDB header: allergen Chain: A; PDB Molecule: cysteine protease; PDBTitle: 2.70 a crystal structure of the amb a 11 cysteine protease, a major2 ragweed pollen allergen, in its proform
3	c3qj3B	 Alignment		100.0	35	PDB header: hydrolase Chain: B; PDB Molecule: cathepsin I-like protein; PDBTitle: structure of digestive procathepsin I2 proteinase from tenebrio2 molitor larval midgut
4	c2c0yA	 Alignment		100.0	35	PDB header: hydrolase Chain: A; PDB Molecule: procathepsin s; PDBTitle: the crystal structure of a cys25ala mutant of human2 procathepsin s
5	c3qt4A	 Alignment		100.0	37	PDB header: hydrolase Chain: A; PDB Molecule: cathepsin-I-like midgut cysteine proteinase; PDBTitle: structure of digestive procathepsin I 3 of tenebrio molitor larval2 midgut
6	c5ef4A	 Alignment		100.0	33	PDB header: allergen Chain: A; PDB Molecule: cysteine protease; PDBTitle: 2.05 a crystal structure of the amb a 11 cysteine protease, a major2 ragweed pollen allergen, in its proform
7	c2o6xA	 Alignment		100.0	36	PDB header: hydrolase Chain: A; PDB Molecule: secreted cathepsin I 1; PDBTitle: crystal structure of procathepsin I1 from fasciola hepatica
8	d1pcia	 Alignment		100.0	33	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
9	d1cs8a	 Alignment		100.0	36	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
10	c3tnxA	 Alignment		100.0	33	PDB header: hydrolase Chain: A; PDB Molecule: papain; PDBTitle: structure of the precursor of a thermostable variant of papain at 2.62 angstrom resolution
11	c5jt8B	 Alignment		100.0	29	PDB header: allergen Chain: B; PDB Molecule: blo t 1 allergen; PDBTitle: structural basis for the limited antibody cross reactivity between the2 mite allergens blo t 1 and der p 1

12	dlxkga1	Alignment		100.0	25	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
13	c5fpwC_	Alignment		100.0	23	PDB header: hydrolase Chain: C: PDB Molecule: pro cathepsin b s9; PDBTitle: procathepsin b s9 from trypanosoma congolense
14	d3pbha_	Alignment		100.0	28	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
15	c4hwyA_	Alignment		100.0	26	PDB header: hydrolase Chain: A: PDB Molecule: cysteine peptidase c (cpc); PDBTitle: trypanosoma brucei procathepsin b solved from 40 fs free-electron2 laser pulse data by serial femtosecond x-ray crystallography
16	d1mira_	Alignment		100.0	28	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
17	c1jqpA_	Alignment		100.0	32	PDB header: hydrolase Chain: A: PDB Molecule: dipeptidyl peptidase i; PDBTitle: dipeptidyl peptidase i (cathepsin c), a tetrameric cysteine protease2 of the papain family
18	d1deua_	Alignment		100.0	27	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
19	d1m6da_	Alignment		100.0	53	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
20	d1me4a_	Alignment		100.0	44	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
21	c3bwbD_	Alignment	not modelled	100.0	39	PDB header: hydrolase Chain: D: PDB Molecule: cysteine protease falcipain-3; PDBTitle: crystal structure of falcipain-3 with its inhibitor, k11017
22	d2as8a1	Alignment	not modelled	100.0	26	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
23	d2r6na1	Alignment	not modelled	100.0	40	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
24	d1fh0a_	Alignment	not modelled	100.0	42	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
25	d2h7ja1	Alignment	not modelled	100.0	43	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
26	d1jqpa2	Alignment	not modelled	100.0	33	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
27	d1cqda_	Alignment	not modelled	100.0	40	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
28	c4ci7A_	Alignment	not modelled	100.0	16	PDB header: hydrolase Chain: A: PDB Molecule: cell surface protein (putative cell surface-associated) PDBTitle: the crystal structure of the cysteine protease and lectin-like2 domains of cwp84, a surface layer associated protein of clostridium3 difficile

29	c3hwnC_	Alignment	not modelled	100.0	44	PDB header: hydrolase Chain: C: PDB Molecule: cathepsin I1; PDBTitle: cathepsin I with az13010160
30	dlaeca_	Alignment	not modelled	100.0	38	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
31	dlppoa_	Alignment	not modelled	100.0	37	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
32	dlvsna1	Alignment	not modelled	100.0	39	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
33	c2b1nA_	Alignment	not modelled	100.0	39	PDB header: sugar binding protein Chain: A: PDB Molecule: spe31; PDBTitle: crystal structure of a papain-fold protein without the catalytic2 cysteine from seeds of pachyrhizus erosus
34	d2oula1	Alignment	not modelled	100.0	34	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
35	d2acta_	Alignment	not modelled	100.0	36	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
36	c3f75A_	Alignment	not modelled	100.0	38	PDB header: hydrolase Chain: A: PDB Molecule: cathepsin I protease; PDBTitle: activated toxoplasma gondii cathepsin I (tgcpl) in complex with its2 propeptide
37	c3hbiB_	Alignment	not modelled	100.0	28	PDB header: hydrolase Chain: B: PDB Molecule: cathepsin b-like cysteine protease; PDBTitle: crystal structure of cathepsin b from t. brucei in complex with ca074
38	c3ioqA_	Alignment	not modelled	100.0	39	PDB header: hydrolase Chain: A: PDB Molecule: cms1ms2; PDBTitle: crystal structure of the carica candamarcensis cysteine protease2 cms1ms2 in complex with e-64.
39	c5a24A_	Alignment	not modelled	100.0	38	PDB header: hydrolase Chain: A: PDB Molecule: dionain-1; PDBTitle: crystal structure of dionain-1, the major endopeptidase in2 the venus flytrap digestive juice
40	c2fo5A_	Alignment	not modelled	100.0	41	PDB header: hydrolase/hydrolase inhibitor Chain: A: PDB Molecule: cysteine proteinase ep-b 2; PDBTitle: crystal structure of recombinant barley cysteine endoprotease b2 isoform 2 (ep-b2) in complex with leupeptin
41	d1iwda_	Alignment	not modelled	100.0	35	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
42	d1s4va_	Alignment	not modelled	100.0	42	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
43	d1gece_	Alignment	not modelled	100.0	38	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
44	c4yyqA_	Alignment	not modelled	100.0	35	PDB header: hydrolase Chain: A: PDB Molecule: ficin isoform a; PDBTitle: ficin a
45	d1o0ea_	Alignment	not modelled	100.0	36	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
46	c1nb3D_	Alignment	not modelled	100.0	38	PDB header: hydrolase Chain: D: PDB Molecule: cathepsin h; PDBTitle: crystal structure of stefin a in complex with cathepsin h: n-terminal2 residues of inhibitors can adapt to the active sites of endo-and3 exopeptidases
47	d1khqa_	Alignment	not modelled	100.0	38	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
48	c3qsdA_	Alignment	not modelled	100.0	30	PDB header: hydrolase/hydrolase inhibitor Chain: A: PDB Molecule: cathepsin b-like peptidase (c01 family); PDBTitle: structure of cathepsin b1 from schistosoma mansoni in complex with2 ca074 inhibitor
49	c3u8eA_	Alignment	not modelled	100.0	37	PDB header: hydrolase Chain: A: PDB Molecule: papain-like cysteine protease; PDBTitle: crystal structure of cysteine protease from bulbs of crocus sativus at2 1.3 a resolution
50	c4yywA_	Alignment	not modelled	100.0	34	PDB header: hydrolase Chain: A: PDB Molecule: ficin isoform d; PDBTitle: ficin d2
51	d1gmya_	Alignment	not modelled	100.0	32	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
52	d1yala_	Alignment	not modelled	100.0	36	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
53	d2dcca1	Alignment	not modelled	100.0	31	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
54	d1ef7a_	Alignment	not modelled	100.0	30	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
55	d1thea_	Alignment	not modelled	100.0	31	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like

56	c2bdzA_	Alignment	not modelled	100.0	37	PDB header: hydrolase Chain: A: PDB Molecule: mexicain; PDBTitle: mexicain from jacaratia mexicana
57	c3ch3X_	Alignment	not modelled	100.0	22	PDB header: hydrolase Chain: X: PDB Molecule: serine-repeat antigen protein; PDBTitle: crystal structure analysis of sera5e from plasmodium falciparum
58	c3oisA_	Alignment	not modelled	100.0	21	PDB header: hydrolase Chain: A: PDB Molecule: cysteine protease; PDBTitle: crystal structure xylellain, a cysteine protease from xylella2 fastidiosa
59	c3pw3E_	Alignment	not modelled	100.0	26	PDB header: hydrolase Chain: E: PDB Molecule: aminopeptidase c; PDBTitle: crystal structure of a cysteine protease (bdi_2249) from2 parabacteroides distasonis atcc 8503 at 2.23 a resolution
60	d3gcba_	Alignment	not modelled	100.0	21	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
61	c1csbE_	Alignment	not modelled	100.0	31	PDB header: hydrolase/hydrolase inhibitor Chain: E: PDB Molecule: cathepsin b heavy chain; PDBTitle: crystal structure of cathepsin b inhibited with ca030 at 2.1 angstroms2 resolution: a basis for the design of specific epoxysuccinyl3 inhibitors
62	c1icfA_	Alignment	not modelled	100.0	44	PDB header: hydrolase Chain: A: PDB Molecule: protein (cathepsin l: heavy chain); PDBTitle: crystal structure of mhc class ii associated p41 ii fragment in2 complex with cathepsin l
63	c1k3bB_	Alignment	not modelled	100.0	29	PDB header: hydrolase Chain: B: PDB Molecule: dipeptidyl-peptidase i light chain; PDBTitle: crystal structure of human dipeptidyl peptidase i (cathepsin c):2 exclusion domain added to an endopeptidase framework creates the3 machine for activation of granular serine proteases
64	c2djgC_	Alignment	not modelled	99.9	41	PDB header: hydrolase Chain: C: PDB Molecule: dipeptidyl-peptidase 1; PDBTitle: re-determination of the native structure of human dipeptidyl peptidase2 i (cathepsin c)
65	c4k7cA_	Alignment	not modelled	99.8	23	PDB header: hydrolase Chain: A: PDB Molecule: aminopeptidase c; PDBTitle: crystal structure of pepw from lactobacillus rhamnosis hn001 (dr20)2 determined as the selenomet derivative
66	d2cb5a_	Alignment	not modelled	99.8	21	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
67	c3f75P_	Alignment	not modelled	99.7	31	PDB header: hydrolase Chain: P: PDB Molecule: cathepsin l propeptide; PDBTitle: activated toxoplasma gondii cathepsin l (tgclp) in complex with its2 propeptide
68	c2l95A_	Alignment	not modelled	99.7	27	PDB header: hydrolase Chain: A: PDB Molecule: crammer; PDBTitle: solution structure of cytotoxic t-lymphocyte antigen-2(ctla protein).2 crammer at ph 6.0
69	c1icfB_	Alignment	not modelled	99.6	41	PDB header: hydrolase Chain: B: PDB Molecule: protein (cathepsin l: light chain); PDBTitle: crystal structure of mhc class ii associated p41 ii fragment in2 complex with cathepsin l
70	c1hucC_	Alignment	not modelled	99.5	33	PDB header: thiol protease Chain: C: PDB Molecule: cathepsin b; PDBTitle: the refined 2.15 angstroms x-ray crystal structure of human2 liver cathepsin b: the structural basis for its specificity
71	d1cv8a_	Alignment	not modelled	97.3	14	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
72	c1y4hA_	Alignment	not modelled	96.8	18	PDB header: hydrolase/hydrolase inhibitor Chain: A: PDB Molecule: cysteine protease; PDBTitle: wild type staphopain-staphostatin complex
73	c1x9yD_	Alignment	not modelled	95.2	18	PDB header: hydrolase Chain: D: PDB Molecule: cysteine proteinase; PDBTitle: the prostaphopain b structure
74	d1pxva_	Alignment	not modelled	95.1	15	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
75	c3ervA_	Alignment	not modelled	94.9	14	PDB header: structural genomics, unknown function Chain: A: PDB Molecule: putative c39-like peptidase; PDBTitle: crystal structure of an putative c39-like peptidase from bacillus2 anthracis
76	d1dkia_	Alignment	not modelled	93.9	29	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
77	c3bb7A_	Alignment	not modelled	92.4	32	PDB header: hydrolase Chain: A: PDB Molecule: interpain a; PDBTitle: structure of prevotella intermedia prointerpain a fragment 39-3592 (mutant c154a)
78	c2jtcA_	Alignment	not modelled	92.1	29	PDB header: hydrolase Chain: A: PDB Molecule: streptopain; PDBTitle: 3d structure and backbone dynamics of spe b
79	c3bbaB_	Alignment	not modelled	91.7	25	PDB header: hydrolase Chain: B: PDB Molecule: interpain a; PDBTitle: structure of active wild-type prevotella intermedia interpain a2 cysteine protease
80	d1pvja_	Alignment	not modelled	90.9	29	Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Papain-like
81	c6bdtC_	Alignment	not modelled	59.9	19	PDB header: hydrolase Chain: C: PDB Molecule: calpain-3; PDBTitle: crystal structure of human calpain-3 protease core

					mutant-c129s
82	d1zcma1	Alignment	not modelled	56.0	30 Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Calpain large subunit, catalytic domain (domain II)
83	d2r9fa1	Alignment	not modelled	52.3	27 Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Calpain large subunit, catalytic domain (domain II)
84	c3g14B_	Alignment	not modelled	50.3	20 PDB header: oxidoreductase Chain: 8: PDB Molecule: nitroreductase family protein; PDBTitle: crystal structure of nitroreductase family protein (yp_877874.1) from2 clostridium novyi nt at 1.75 a resolution
85	d1mdwa_	Alignment	not modelled	47.4	34 Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Calpain large subunit, catalytic domain (domain II)
86	c1zivA_	Alignment	not modelled	43.3	32 PDB header: hydrolase Chain: A: PDB Molecule: calpain 9; PDBTitle: catalytic domain of human calpain-9
87	d1ziva1	Alignment	not modelled	43.3	32 Fold: Cysteine proteinases Superfamily: Cysteine proteinases Family: Calpain large subunit, catalytic domain (domain II)
88	d2nn6h3	Alignment	not modelled	41.2	14 Fold: Eukaryotic type KH-domain (KH-domain type I) Superfamily: Eukaryotic type KH-domain (KH-domain type I) Family: Eukaryotic type KH-domain (KH-domain type I)
89	d1is3a_	Alignment	not modelled	37.5	23 Fold: Concanavalin A-like lectins/glucanases Superfamily: Concanavalin A-like lectins/glucanases Family: Galectin (animal S-lectin)
90	d1xnda_	Alignment	not modelled	36.6	26 Fold: Concanavalin A-like lectins/glucanases Superfamily: Concanavalin A-like lectins/glucanases Family: Xylanase/endoglucanase 11/12
91	c3sumA_	Alignment	not modelled	31.0	33 PDB header: unknown function Chain: A: PDB Molecule: cerato-platanin-like protein; PDBTitle: crystal structure of cerato-platanin 5 from m. perniciosa (mpcp5)
92	d1ynaa_	Alignment	not modelled	30.7	21 Fold: Concanavalin A-like lectins/glucanases Superfamily: Concanavalin A-like lectins/glucanases Family: Xylanase/endoglucanase 11/12
93	d1ulea_	Alignment	not modelled	30.0	38 Fold: Concanavalin A-like lectins/glucanases Superfamily: Concanavalin A-like lectins/glucanases Family: Galectin (animal S-lectin)
94	c6b8hd_	Alignment	not modelled	24.6	13 PDB header: membrane protein Chain: D: PDB Molecule: atp synthase subunit beta, mitochondrial; PDBTitle: mosaic model of yeast mitochondrial atp synthase monomer
95	d1n62a2	Alignment	not modelled	22.5	37 Fold: beta-Grasp (ubiquitin-like) Superfamily: 2Fe-2S ferredoxin-like Family: 2Fe-2S ferredoxin domains from multidomain proteins
96	d1t3qa2	Alignment	not modelled	22.2	37 Fold: beta-Grasp (ubiquitin-like) Superfamily: 2Fe-2S ferredoxin-like Family: 2Fe-2S ferredoxin domains from multidomain proteins
97	d1jroa2	Alignment	not modelled	21.9	44 Fold: beta-Grasp (ubiquitin-like) Superfamily: 2Fe-2S ferredoxin-like Family: 2Fe-2S ferredoxin domains from multidomain proteins
98	c3m3gA_	Alignment	not modelled	20.1	33 PDB header: polysaccharide-binding protein Chain: A: PDB Molecule: repl1 protein; PDBTitle: crystal structure of sm1, an elicitor of plant defence responses from2 trichoderma virens.
99	c3g3oA_	Alignment	not modelled	19.3	8 PDB header: biosynthetic protein Chain: A: PDB Molecule: vacuolar transporter chaperone 2; PDBTitle: crystal structure of the cytoplasmic tunnel domain in yeast2 vtc2p

Phyre²

Domain analysis

Unique Job ID 93655db4c2450b0b

Rank	Aligned region
1	d7pcka_
2	c5egwA_
3	c3qj3B_
4	c2c0vA_
5	c3qt4A_
6	c5ef4A_
7	c2o6xA_
8	d1pcia_
9	d1cs8a_
10	c3tnxA_
11	c5jt8B_
12	d1xkga1
13	c5fpwC_
14	d3pbha_
15	c4hwyA_
16	d1mira_
17	c1jqpA_
18	d1deua_
19	d1m6da_
20	d1me4a_
21	
22	
23	
24	
25	
26	
27	
28	
29	
30	
31	
32	
33	
34	
35	
36	
37	
38	
39	
40	
41	
42	
43	
44	
45	
46	
47	
48	
49	
50	
51	
52	
53	
54	
55	
56	
57	
58	



S3.1 Chapter 3 Supplementary File 1

Statistically significant differences in NanoDrop 2000™ and Qubit™ DNA concentration, and 260/230 nm absorbance ratio using the 11 DNA extraction methods according to the Dunn's Multiple Comparison Test.

Method Comparison				
NanoDrop 2000™ DNA Concentration		Z	P. unadj	P. adj
AccM	CheX	-4.03309	5.50E-05	2.86E-03
CheX	CTAB	4.665147	3.08E-06	1.67E-04
CTAB	EznF	-3.64935	2.63E-04	1.26E-02
CTAB	Schi	-3.83746	1.24E-04	6.22E-03
CheX	Schi-LE	3.574104	3.51E-04	1.65E-02
CTAB	SDS	-3.49886	4.67E-04	2.15E-02
CheX	WizM	4.9586	7.10E-07	3.91E-05
EznF	WizM	3.942802	8.05E-05	4.11E-03
Schi	WizM	4.130912	3.61E-05	1.92E-03
SDS	WizM	3.792313	1.49E-04	7.31E-03
Qubit™ DNA Concentration				
AccM	CheX	-3.56893	3.58E-04	0.012187
AccW	CheX	-3.92687	8.61E-05	0.003098
AccM	Schi	-3.50576	4.55E-04	0.015025
AccW	Schi	-3.86371	1.12E-04	0.003909
CheX	Schi-LE	3.386779	7.07E-04	0.02263
Schi	Schi-LE	3.326552	8.79E-04	0.027258
A260/230				
Optimal*	Schi-LE	3.898243	9.69E-05	0.006007
EznF	SDS	4.029334	5.59E-05	0.00358
IsoG	SDS	4.029334	5.59E-05	0.003524
Optimal*	SDS	4.70549	2.53E-06	0.000167
EznF	WizM	3.725754	1.95E-04	0.011879
IsoG	WizM	3.725754	1.95E-04	0.011684
Optimal*	WizM	4.40191	1.07E-05	0.000697

* Optimal ratio for pure DNA using A260/230 measurement is 2.0.

Z: Values for the Z test statistic for each comparison.

P. unadj: Unadjusted p-values for each comparison.

P. adj: Adjusted p-values for each comparison.

S3.2 Chapter 3 Supplementary File 2

DNA extraction methods

AccM

AccuPrep Genomic DNA Extraction – Mammalian Tissue

Disrupt or homogenise the sample into a clean 1.5 ml tube and add 200 µL of Tissue Lysis buffer. Add 20 µL of Proteinase K, mix by vortex mixer, and incubate at 60°C for 1 hour or until the tissue is completely lysed. Briefly spin down to remove drops from inside the lid and sides. Add 200 µL Binding buffer and immediately vortex. Incubate at 60°C for 10 mins. Add 100 µL isopropanol and mix well by pipetting. Carefully transfer the lysate into the binding column tube. Centrifuge at 8,000 rpm for 1 min. Transfer the binding column tube to a new 2 ml tube. Add 500 µL Washing buffer 1 and centrifuge at 8,000 rpm for 1 min. Dispose of flow-through. Add 500 µL of Washing buffer 2 and centrifuge at 8,000 rpm for 2 mins. Transfer the binding column tube to a clean 1.5 ml tube. Add 30 µL of Elution buffer and incubate at room temperature for 1 min. Centrifuge at 8,000 rpm for 1 min.

AccW

AccuPrep Genomic DNA Extraction – Whole Blood, Buffy Coat and Cultured Cells

Add 20 µL of Proteinase K to a clean 1.5 ml tube. Add sample, 200 µL of Binding buffer and vortex immediately. Incubate at 60°C for 10 mins. Add 100 µL isopropanol and mix well by pipetting. Carefully transfer the lysate into the binding column tube. Centrifuge at 8,000 rpm for 1 min. Transfer the binding column tube to a new 2 ml tube. Add 500 µL Washing buffer 1 and centrifuge at 8,000 rpm for 1 min. Dispose of flow-through. Add 500 µL of Washing buffer 2 and centrifuge at 8,000 rpm for 2 mins. Transfer the binding column tube to a clean 1.5 ml tube. Add 30 µL of Elution buffer and incubate at room temperature for 1 min. Centrifuge at 8,000 rpm for 1 min.

CheX

Chelex-100

Aliquot 300 µL 5% Chelex solution into 1.5 ml tubes. Under dissection microscope, place sample on a slide and crush with another clean microscope slide. Use a needle-stick to collect crushed sample and place into tube with 5% Chelex. Incubate tubes at 56°C for 15 mins, vortex thoroughly then incubate at 100 °C for 8 mins. Vortex tubes and centrifuge at 15,000 rpm for 5-10 mins. Transfer 200 µL of supernatant into new 1.5 ml tube.

CTAB

CTAB DNA Extraction

Add 500 µL 2X CTAB buffer (100 mM Tris-HCl pH 8, 1.4 M NaCl, 20 mM EDTA, 2% CTAB, 0.02 g / 1 ml PVP-40) to 2 ml tube with sample. Using pellet pestle, grind sample. Perform short grinds (30 seconds max.) followed by cooling on ice to avoid heat damage to DNA. Add 500 µL 2X CTAB buffer heated to 65°C and incubate at 65°C for 60 mins. Add 750 µL SEVAG (24:1 chloroform:isoamyl alcohol) and mix gently on orbital shaker for up to 1 hour. Centrifuge at 13,000 rpm for 10 mins. Transfer top aqueous phase to a new 1.5 ml tube. Add 750 µL SEVAG and mix gently on orbital shaker for up to 1 hour. Centrifuge at 13,000 rpm for 10 mins. Transfer top aqueous phase to a new 1.5 ml tube. Add 2/3 volume of -20°C isopropanol and mix gently. Centrifuge at 13,000 rpm for 10 mins. Pour off liquid in waste container in fume hood. Add 750 µL of ice-cold 70% ethanol, shake well, and wash for 60 mins. Centrifuge at 13,000 rpm for 5 mins. Pour off liquid and drain upside-down for 5-10 mins. If ethanol still present in tubes do an extra spin in centrifuge and pipette ethanol out carefully. Resuspend DNA in 100 µL TE buffer pH 8 (10 mM Tris-HCl, 1 mM EDTA). Add 1 µL 100 mg/ml RnaseA and vortex. Incubate at 37°C for 20 mins. Add 100 µL of cold 7.5 M ammonium acetate and 750 µL 100% ethanol. Incubate on ice for 30 mins. Centrifuge at 13,000 rpm for 10 mins. Pour off liquid in waste container in fume hood. Add 500 µL ice-cold 70% ethanol and shake well. Wash for 30 mins. Centrifuge at 13,000 rpm for 5 mins. Pour off supernatant and air-dry upside-down for 5-10 mins. Dissolve DNA in 50 µL dH₂O.

EznF

E.Z.N.A.® Forensic DNA Kit – Standard Protocol

Place sample into 1.5 ml tube, add 200 µL TL Buffer and vortex. Incubate at 55°C for 15 mins. Vortex every 2 mins. Add 25 µL OB Protease Solution and vortex. Incubate at 60°C for 45 mins with occasional mixing. Centrifuge at maximum speed to collect any sample adhering to walls. Add 225 µL BL Buffer and vortex. Incubate at 60°C for 10 mins. Centrifuge at max. speed to collect any sample adhering to walls. Add 300 µL 100% isopropanol and vortex. Centrifuge at max. speed to collect any sample adhering to walls. Insert column to 2 ml collection tube. Add 100 µL 3 M NaOH to column and incubate at room temperature for 4 mins. Centrifuge at max. speed for 30 secs. Discard filtrate and reuse collection tube. Transfer sample to column and centrifuge at max. speed for 1 min. Discard filtrate and collection tube. Transfer column to new 2 ml collection tube, add 500 µL HBC Buffer and centrifuge at max. speed for 1 min. Discard filtrate and collection tube. Transfer column to a new 2 ml collection tube, add 700 µL DNA Wash Buffer and centrifuge at max. speed for 1 min. Discard the filtrate and reuse the collection tube. Do a second wash by adding 700 µL DNA Wash Buffer and centrifuge at max. speed for 2 mins. Place the column into a clean 1.5 ml tube and add 30 µL Elution Buffer heated to 70°C. Incubate at room temperature for 3 minutes and then centrifuge at max. speed for 1 min.

IsoG

ISOLATE II Genomic DNA Kit

Add 180 µL Lysis Buffer GL and sample to 1.5 ml tube. Add 25 µL Proteinase K and vortex. Incubate at 56°C for 1-3 hours, until completely lysed, vortex occasionally. Add 200 µL Lysis Buffer G3 and vortex vigorously. Incubate at 70°C for 10 mins. Vortex briefly and add 210 µL 100% ethanol, vortex again. Place spin column in 2 ml collection tube and load sample to the column. Centrifuge at 11,000 x *g* for 1 min. Discard flow-through and reuse collection tube. Add 500 µL Wash Buffer GW1 to column and centrifuge for 1 min at 11,000 x *g*. Discard the flow-through and reuse the collection tube. Add 600 µL Wash Buffer GW2 and centrifuge for 1 min at 11,000 x *g*. Discard the flow-through and reuse collection tube. Centrifuge for 1 min at 11,000 x *g* to remove residual ethanol. Place column in a clean 1.5 ml tube. Add 30 µL Elution Buffer G preheated to 70°C to column. Incubate at room temperature for 1 min. Centrifuge for 1 min at 11,000 x *g*.

Schi

Schistosoma sp. DNA Extraction

Mix Homogenisation Buffer (100 mM NaCl, 200 mM sucrose, 10 mM EDTA, 30 mM Tris pH 8) and Lysis Buffer (250 mM EDTA pH 8, 2.5% SDS, 500 mM Tris pH 9.2) at 4:1 to create fresh Grinding Buffer (GB). Grind 1 nematode in 25 µL GB in 1.5 ml tube. Rinse the pestle with 25 µL of fresh GB and add wash to sample. Incubate at 65°C for 30 mins. Add 8 M ammonium acetate to a final concentration of 1 M. Incubate on wet ice for 30-60 mins. Centrifuge at 13,000 rpm for 10 mins. Transfer supernatant to a clean 1.5 ml tube. Add 100 µL 100% ethanol, mix, and incubate at room temperature for 5 mins. Centrifuge for 15 mins at 13,000 rpm at 4°C. Remove ethanol carefully. Wash barely visible pellet with 100% ethanol once, remove ethanol carefully and let tube air-dry but not over dry. Add 30 µL TE buffer (10 mM Tris, 1 mM EDTA, pH 8).

SDS

Sodium Dodecyl Sulphate (SDS) DNA Extraction

Add 150 µL of DNA extraction buffer (200 mM Tris pH 8, 250 mM NaCl, 125 mM EDTA, 0.5% SDS) and sample to 1.5 ml tube, and crush with a sterile pellet pestle. Add 150 µL of DNA extraction buffer to was pellet pestle into the tube. Vortex for 30 sec. Freeze at -20°C for 10 mins. Incubate at 70°C for 10 mins, vortex at 5 mins. Incubate at 4°C for 5 mins. Centrifuge at 13,000 rpm for 10 mins at 4°C. Transfer the supernatant to a new 1.5 ml tube containing 300 µL ice-cold isopropanol. Mix by pipetting. Centrifuge at 13,000 rpm for 10 mins at 4°C. Remove and discard the supernatant without disturbing the pellet. Wash the DNA pellet with 300 µL of ice-cold ethanol and mix by pipetting. Centrifuge at 13,000 rpm for 10 mins at 4°C. Remove and discard ethanol supernatant

without disturbing the pellet and air-dry tube until all traces of ethanol have evaporated. Resuspend DNA pellet in 30 μ L dH₂O.

WizM

Wizard Genomic DNA Purification Kit – Animal Tissue (Mouse Tail)

Add 120 μ L of 0.5 M EDTA (pH 8) and 500 μ L of Nuclei Lysis Solution to a 1.5 ml tube and chill on ice. Add sample to a new 1.5 ml tube and add 600 μ L of EDTA/Nuclei Lysis Solution. Add 17.5 μ L of 20 mg/ml Proteinase K. Incubate at 55°C for 3 hours, vortex the sample once per hour. Add 3 μ L of RNase Solution and mix by inversion 2-5 times. Incubate at 37°C for 30 mins. Cool to room temperature for 5 mins. Centrifuge for 4 mins at 13,000 x *g*. Transfer the supernatant to a clean 1.5 ml tube containing 600 μ L isopropanol. Mix by inversion. Centrifuge for 1 min at 13,000 x *g*. Remove and discard of the supernatant carefully. Air-dry the pellet for 10-15 mins. Add 30 μ L Rehydration Solution and incubate at 4°C overnight.

WizP

Wizard Genomic DNA Purification Kit – Plant Tissue

Add sample to 600 μ L of Nuclei Lysis Solution and vortex for 1-3 secs. Incubate at 65°C for 15 mins. Add 3 μ L RNase Solution and mix by inversion. Incubate at 37°C for 15 mins. Cool to room temperature for 5 mins. Add 200 μ L of Protein Precipitation Solution and vortex vigorously for 20 seconds. Centrifuge for 3 mins at 13,000 x *g*. Transfer supernatant to a clean 1.5 ml tube containing 600 μ L of isopropanol. Mix by inversion. Centrifuge for 1 min at 13,000 x *g*. Remove and dispose of supernatant. Add 600 μ L of 70% ethanol and invert gently. Centrifuge at 13,000 x *g* for 1 min. Remove and discard of the supernatant carefully. Air-dry the pellet for 10-15 mins. Add 30 μ L Rehydration Solution and incubate at 4°C overnight.

-LE

Larval exsheathment

Nematodes washed twice by centrifugation in distilled water at 300 x *g* for 5 mins. Discard supernatant between washes. Pellet sample by centrifugation at 13,000 x *g* for 30 secs. Resuspend in 400 μ L distilled water. Add 400 μ L 1% sodium hypochlorite, to give a final conc. of 1.5%. Incubate at 40°C for 10 mins. Remove supernatant and wash in distilled water. Centrifuge at 13,000 x *g* for 30 secs. Discard supernatant and wash again in distilled water. Centrifuged at 13,000 x *g* for 30 secs. Discard supernatant and resuspend in 400 μ L distilled water. Incubate at 40°C in a 10% CO₂ atmosphere for 30 mins. Store at 4°C until required for DNA extraction.