



University  
of Glasgow

Johnston, Stephanie Lauren (2012) *Identification of multigene cysteine protease gene families in Haemonchus contortus and analysis of gut gene expression*. PhD thesis.

<http://theses.gla.ac.uk/3735/>

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author

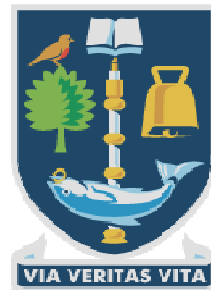
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten:Theses  
<http://theses.gla.ac.uk/>  
theses@ gla.ac.uk

**Identification of multigene cysteine protease gene families in *Haemonchus contortus* and analysis of gut gene expression**

**STEPHANIE LAUREN JOHNSTON, BVMS**



**UNIVERSITY  
of  
GLASGOW**

**Submitted in fulfilment of the requirements for the degree of  
Doctor of Philosophy**

**College of Medical, Veterinary and Life Sciences  
School of Veterinary Medicine  
Division of Infection and Immunity  
University of Glasgow  
2012**

**© Stephanie Johnston, 2012**

## **Author's Declaration**

I hereby declare that the work presented in this thesis is entirely my own original work and carried out under the supervision of Dr Collette Britton. Where other sources of information have been used, they have been acknowledged.

Stephanie Johnston

September 2012

## Abstract

*Haemonchus contortus* is a blood-feeding Strongylid parasite that is economically significant worldwide. Due to the increasing problem of anthelmintic resistance, alternative approaches are urgently required for parasitic nematode control. *H. contortus* cathepsin B gut cysteine proteases have received attention as potential vaccine candidates because of their proposed role in blood feeding. The increasing amount of *H. contortus* genome information has now enabled detailed identification and annotation of cathepsin B protease gene families. In this study *H. contortus* BAC 18f22 was annotated and found to encode eight tandemly arranged cysteine proteases related to the previously identified AC family, but with six novel genes identified. Annotation of supercontig and scaffold sequence identified many more members of the HmCP and GCP-7 cathepsin B families. In total this work has shown that the *H. contortus* genome encodes at least 41 cathepsin B protease genes, more than in other nematodes, to date. In contrast, *Hc-cpr-6* is present as a single copy gene that is highly conserved in a number of species, suggesting an important conserved function.

Further work examined regulation of gut gene expression in *H. contortus*, in particular the *H. contortus* ELT-2 GATA transcription factor (TF), as it has been shown to be the major TF in *C. elegans* controlling gut gene expression. A high throughput assay was developed and used to screen an integrated *C. elegans* worm strain expressing GFP in the gut and hypodermis (*Ce-cpl-1::gfp*) against 594 chemical compounds. Compounds were identified that specifically cause a decrease in gut GFP expression, affect larval development and show a degree of lethality. Further work on two of the compounds identified an embryonic effect, with a significant decrease in number of progeny. To conclude, this thesis identified a number of novel cathepsin B genes as well as two compounds potentially interfering with TF activity and gut gene expression which may be of use as novel anthelmintics.

# Table of Contents

|  |           |
|--|-----------|
| Author's Declaration .....   | ii        |
| Abstract .....   | iii       |
| Table of Contents.....   | iv        |
| List of Figures .....  | ix        |
| List of Tables.....  | xi        |
| Acknowledgements .....   | xii       |
| List of Abbreviations and Symbols.....                                       | xiii      |
| <b>Chapter 1 Introduction .....</b>  | <b>15</b> |
| 1.1 Parasitic nematode infections .....                                      | 16        |
| 1.1.1 <i>Haemonchus contortus</i> .....                                      | 16        |
| 1.1.2 <i>H. contortus</i> life cycle .....                                   | 17        |
| 1.1.3 The use of anthelmintic drugs to treat parasitic infection.....        | 19        |
| 1.1.4 Anthelmintic resistance .....  | 21        |
| 1.1.5 Alternative approaches for <i>H. contortus</i> control .....           | 23        |
| 1.1.6 Nematode genome sequencing .....                                       | 25        |
| 1.1.7 <i>Caenorhabditis elegans</i> as a model for parasitic nematodes ..... | 27        |
| 1.2 Gene regulation .....  | 28        |
| 1.2.1 Gene regulation in nematodes .....                                     | 28        |
| 1.2.2 Regulation of gut expressed genes in nematodes.....                    | 30        |
| 1.3 Aims and Objectives .....  | 31        |
| <b>Chapter 2 Materials and methods.....</b>                                  | <b>32</b> |
| 2.1 <i>C. elegans</i> methods .....  | 33        |
| 2.1.1 Culture and maintenance .....  | 33        |
| 2.1.2 Transformation of <i>C. elegans</i> by microinjection.....             | 34        |
| 2.1.3 Staining and viewing <i>C. elegans</i> worms .....                     | 35        |
| 2.1.4 Transgene integration .....  | 35        |
| 2.1.5 <i>C. elegans</i> RNA interference (RNAi) .....                        | 37        |
| 2.1.6 Antibody localisation methods .....                                    | 38        |
| 2.1.6.1 Generation of an anti-peptide antibody.....                          | 38        |
| 2.1.6.2 Preparation of protein extracts.....                                 | 38        |
| 2.1.6.3 Protein separation by polyacrylamide gel electrophoresis ....        | 39        |
| 2.1.6.4 Western Blotting of Ce-CPR-6.....                                    | 39        |
| 2.1.6.5 Freeze-crack method for worm immunofluorescence .....                | 40        |

|   |  |    |
|---|--|----|
| 2.1.6.6   | CPR-6 antibody staining method.....  | 40 |
| 2.1.7   | <i>C. elegans</i> drug screening .....   | 40 |
| 2.2   | <i>H. contortus</i> methods .....  | 41 |
| 2.2.1   | <i>H. contortus</i> larval culture.....  | 41 |
| 2.2.2   | <i>H. contortus</i> drug screening .....   | 41 |
| 2.3   | Molecular biology methods .....  | 42 |
| 2.3.1   | Polymerase Chain Reaction (PCR).....   | 42 |
| 2.3.1.1   | Standard PCR protocol .....  | 42 |
| 2.3.1.2   | Proof-reading PCR protocol for gene promoter<br>amplification.....                                   | 42 |
| 2.3.1.3   | Rapid Amplification of cDNA Ends (RACE).....   | 42 |
| 2.3.1.4   | Worm lysis PCR.....  | 43 |
| 2.3.1.5   | Reverse Transcription PCR.....   | 44 |
| 2.3.2   | Agarose gel electrophoresis.....   | 44 |
| 2.3.3   | Purification of PCR products.....  | 44 |
| 2.3.4   | Restriction enzyme analysis.....   | 45 |
| 2.3.5   | Cloning into pCR 2.1-TOPO.....   | 45 |
| 2.3.6   | Cloning into Fire lab vectors.....   | 45 |
| 2.3.7   | Identification of positive cultures.....   | 46 |
| 2.3.8   | DNA purification .....   | 46 |
| 2.3.9   | Preparation of compounds for drug screening .....  | 46 |
| 2.4   | Bioinformatic methods .....  | 47 |
| 2.4.1   | Software and databases used .....  | 47 |
| 2.4.2   | DNA primers for PCR .....  | 47 |
| 2.4.3   | DNA sequencing.....  | 47 |
| 2.4.4   | Signal peptide cleavage identification .....   | 48 |
| 2.4.5   | Computational gel analysis for expression levels.....  | 48 |
| <b>Chapter 3 Annotation and characterisation of a cathepsin B</b>                   |  |    |
| <b>cysteine protease gene family present on <i>H. contortus</i> BAC 18f22 .....</b> |  |    |
| 3.1   | Introduction.....  | 50 |
| 3.2   | Results .....  | 52 |
| 3.2.1   | Protease genes on <i>H. contortus</i> BAC 18f22 .....  | 52 |
| 3.2.1.1   | Annotation of genes on <i>H. contortus</i> BAC 18f22 .....   | 52 |
| 3.2.1.2   | Identification of the 5' end of the cysteine protease genes<br>on <i>H. contortus</i> BAC 18f22..... | 55 |

|  |   |            |
|--|---|------------|
| 3.2.1.3  | Intronic analysis of the <i>H. contortus</i> genes on BAC 18f22 .....                                       | 57         |
| 3.2.1.4  | Expression patterns of the BAC genes .....  | 62         |
| 3.2.1.5  | Identification of a complete <i>sj08</i> sequence .....   | 65         |
| 3.2.2  | Analysis of proteases encoded by <i>H. contortus</i> BAC 18f22 .....  | 66         |
| 3.2.2.1  | Similarity of BAC-encoded proteins to the previously<br>identified <i>H. contortus</i> AC family .....      | 66         |
| 3.2.2.2  | Structure and function of cysteine proteases .....  | 71         |
| 3.2.2.3  | Analysis of synonymous and non-synonymous substitutions<br>in cathepsin B cysteine proteases .....          | 75         |
| 3.2.3  | Naming the genes encoded by BAC 18f22 .....   | 77         |
| 3.3  | Discussion .....  | 78         |
| <b>Chapter 4 Other multigene protease families of <i>H. contortus</i> and<br/>analysis of <i>cpr-6</i>, a unique protease gene .....</b> |   | <b>83</b>  |
| 4.1  | Introduction.....   | 84         |
| 4.2  | Results .....   | 85         |
| 4.2.1  | Multigene families in <i>H. contortus</i> .....   | 85         |
| 4.2.1.1  | <i>H. contortus</i> cysteine protease genes related to the HmCP<br>gene family .....                        | 85         |
| 4.2.1.2  | A <i>H. contortus</i> multigene family related to <i>Hc-gcp-7</i> .....                                     | 97         |
| 4.2.2  | Analysis of <i>H. contortus cpr-6</i> , a unique protease gene.....   | 109        |
| 4.2.2.1  | Completion and annotation of the conserved <i>H. contortus</i><br>cysteine protease gene <i>cpr-6</i> ..... | 109        |
| 4.2.2.2  | CPR-6 conservation in other parasitic nematodes .....   | 110        |
| 4.2.2.3  | <i>cpr-6</i> expression in <i>H. contortus</i> adult and larval stages .....                                | 117        |
| 4.2.2.4  | Attempts to localise CPR-6 using specific antibody.....   | 118        |
| 4.2.2.5  | Temporal and spatial expression of a <i>C. elegans cpr-6</i><br>translational reporter .....                | 119        |
| 4.3  | Discussion .....  | 121        |
| <b>Chapter 5 Regulation of gut gene expression in nematodes .....</b>  |   | <b>125</b> |
| 5.1  | Introduction.....   | 126        |
| 5.2  | Results .....   | 128        |
| 5.2.1  | Bioinformatic analysis of <i>H. contortus</i> promoters .....   | 128        |
| 5.2.1.1  | <i>H. contortus</i> gut gene promoter motifs .....  | 128        |
| 5.2.2  | Practical analysis of gut gene expression .....   | 139        |

|   |   |            |
|---|---|------------|
| 5.2.2.1   | Expression pattern of <i>H. contortus</i> <i>sj04</i> (AC-2) promoter in transgenic <i>C. elegans</i> ..... | 139        |
| 5.2.1.2   | RNAi of <i>Ce-elt-2</i> in <i>C. elegans</i> Hc-AC-2 transgenic worm strain.....                            | 141        |
| 5.2.1.3   | RNAi of <i>Ce-elt-2</i> in <i>C. elegans</i> <i>Ce-cpl-1::gfp</i> transgenic worm strain .....              | 141        |
| 5.2.1.4   | Generating an integrated <i>C. elegans</i> <i>cpl-1::gfp</i> strain as a drug screening tool .....          | 142        |
| 5.3   | Discussion .....  | 144        |
| <b>Chapter 6 Screening for potential inhibitors of gut gene expression in <i>C. elegans</i></b> ..... |   | <b>147</b> |
| 6.1   | Introduction.....   | 148        |
| 6.2   | Results .....   | 152        |
| 6.2.1   | Determining the conditions required for screening <i>C. elegans</i> worms .....                             | 152        |
| 6.2.2   | Determining the optimum OP50 concentration for compound screening .....                                     | 153        |
| 6.2.3   | A test screen using <i>C. elegans</i> to identify optimum conditions ...                                    | 154        |
| 6.2.4   | Determining the number of CLB01 worms required per sample well.....   | 158        |
| 6.2.5   | Determining the concentration of DMSO acceptable for screening .....  | 161        |
| 6.2.6   | ELT-2 RNAi as a control to confirm that a decrease in fluorescence will be measurable.....                  | 162        |
| 6.2.7   | Compound selection for use in <i>C. elegans</i> .....   | 164        |
| 6.2.8   | Initial compound screen and analysis of effects .....   | 165        |
| 6.2.9   | Large scale screening of compounds in <i>C. elegans</i> .....   | 168        |
| 6.2.10  | Searching for compounds similar to those of interest .....  | 169        |
| 6.2.11  | Creating a dose response curve for compounds of interest .....  | 170        |
| 6.2.12  | Screening and analysis of compounds of interest.....  | 170        |
| 6.2.13  | Screening of compounds using <i>H. contortus</i> L3 larvae.....   | 176        |
| 6.2.14  | Further analysis of two of the Pfizer compounds.....  | 179        |
| 6.3   | Discussion .....  | 186        |
| <b>Chapter 7 General discussion</b> .....   |   | <b>193</b> |
| <b>References</b> .....   |   | <b>200</b> |



|  |            |
|--|------------|
| <b>Appendices.....</b>                       | <b>215</b> |
| Appendix 1: Common buffers and reagents..... | 216        |
| Appendix 2: Primer sequences.....            | 218        |
| Appendix 3: Additional figures.....          | 220        |

## List of Figures

|   |     |
|---|-----|
| Figure 1.1 Phylogenetic tree depicting the nematode clades .....  | 17  |
| Figure 1.2 <i>H. contortus</i> life cycle .....   | 19  |
| Figure 2.1 Diagrammatic representation of the <i>C. elegans</i> transgenic out-cross protocol .....   | 37  |
| Figure 3.1 Diagrammatic representation of the size and approximate location of genes on <i>H. contortus</i> BAC 18f22 .....                                       | 54  |
| Figure 3.2 Artemis screenshot illustrating the gene structure of the annotated protease genes on <i>H. contortus</i> BAC 18f22 .....                              | 54  |
| Figure 3.3 Confirmation of the 5' UTR of <i>sj02</i> by RT-PCR .....  | 57  |
| Figure 3.4 DNA alignment of the most conserved first introns of certain BAC genes .....   | 59  |
| Figure 3.5 Model indicating the gene structure of the protease genes on BAC 18f22 .....   | 60  |
| Figure 3.6 Artemis screenshot indicating the repeat elements on BAC 18f22 ....  | 61  |
| Figure 3.7 Artemis screenshot indicating transcriptome data associated with two of the BAC genes .....  | 63  |
| Figure 3.8 RT-PCR of BAC protease genes relative to <i>Hc-sod-1</i> .....   | 64  |
| Figure 3.9 SJ08 sequence alignments for the BAC 18f22 and supercontig_0058857 .....   | 66  |
| Figure 3.10 Phylogenetic tree indicating the BAC and AC protease families .....   | 69  |
| Figure 3.11 AC-4 sequence comparison with supercontig_0005737 .....   | 70  |
| Figure 3.12 AC-5 sequence comparison with supercontig_0008756 .....   | 71  |
| Figure 3.13 Amino acid alignment of the <i>H. contortus</i> proteases encoded by BAC 18f22 .....  | 73  |
| Figure 3.14 Graph indicating the rate of synonymous and non-synonymous substitutions in the BAC and AC genes .....  | 76  |
| Figure 4.1 Annotation of the HmCP-related protease genes on supercontig_0059492 .....   | 87  |
| Figure 4.2 Phylogenetic tree indicating the relationship between both the <i>H. contortus</i> cysteine proteases and the related genes on supercontig_0059492 ... | 90  |
| Figure 4.3 Amino acid alignment of the proteases on supercontig_0059492 .....   | 92  |
| Figure 4.4 Annotation of the CBL protease genes on scaffold161 (12062012 file) .....  | 95  |
| Figure 4.5 Phylogenetic tree indicating the relationship between all the proteases on scaffold161 and the HmCP family .....                                       | 96  |
| Figure 4.6 Annotation of the protease genes on supercontig_0041161 and supercontig_0059702 .....  | 98  |
| Figure 4.7 Phylogenetic tree indicating the relationship between all the proteases encoded by supercontig_0041161 and supercontig_0059702 .....                   | 102 |
| Figure 4.8 Amino acid alignment of all the proteases in the GCP-7-like family   | 104 |
| Figure 4.9 Annotation of the protease genes on scaffold306 (12062012 file) ...  | 107 |
| Figure 4.10 Phylogenetic tree indicating the relationship between all the proteases on scaffold306, GCP-7, HmCP1 and HmCP2 .....                                  | 108 |
| Figure 4.11 Diagrammatic representation of the complete <i>H. contortus</i> <i>cpr-6</i> gene .....   | 110 |
| Figure 4.12 Phylogenetic tree indicating the evolution of CPR-6 between a number of parasitic nematodes .....   | 113 |
| Figure 4.13 Amino acid alignment of CPR-6 from <i>H. contortus</i> and other parasitic nematodes .....  | 115 |

|  |     |
|--|-----|
| Figure 4.14 RT-PCR of <i>H. contortus</i> <i>cpr-6</i> expression relative to <i>Hc-sod-1</i> .....  | 118 |
| Figure 4.15 Western blot obtained using the CPR-6 antiserum.....   | 119 |
| Figure 4.16 <i>Ce-cpr-6::gfp</i> translational fusion in transgenic <i>C. elegans</i> worms.   | 120 |
| Figure 5.1 Promoter motif analysis of cathepsin B genes identified on BAC 18f22 .....  | 131 |
| Figure 5.2 Sequence alignment around the E box motifs of the BAC 18f22 genes .....   | 133 |
| Figure 5.3 Promoter analysis of HmCP cathepsin B genes on supercontig_0059492 .....  | 135 |
| Figure 5.4 Promoter analysis of <i>gcp-7</i> -like genes on supercontig_0041161 and supercontig_0059702.....   | 136 |
| Figure 5.5 Sequence alignment around E box motifs of the HmCP family and <i>gcp-7</i> -like genes .....  | 137 |
| Figure 5.6 DNA alignment of the <i>H. contortus</i> and <i>C. elegans</i> <i>cpr-6</i> promoters   | 138 |
| Figure 5.7 Lac-Z expression of <i>C. elegans</i> adult worms transformed with <i>H. contortus</i> AC-2 promoter reporter construct.....  | 140 |
| Figure 5.8 Effect of <i>elt-2</i> RNAi on <i>C. elegans</i> <i>cpl-1::gfp</i> promoter and <i>C. elegans</i> CLB01 worm strain with GFP staining in the gut and hypodermal cells .....     | 143 |
| Figure 6.1 Average fluorescence readings at different OP50 concentrations ...  | 154 |
| Figure 6.2 Diagrammatic representation of the preliminary sample plate .....   | 155 |
| Figure 6.3 Fluorescence readings in different sample conditions in a preliminary screen in the absence of compounds.....   | 158 |
| Figure 6.4 CLB01 worms in sample wells at 2x magnification .....   | 160 |
| Figure 6.5 Percentage death of CLB01 worms at different concentrations of DMSO .....   | 162 |
| Figure 6.6 Level of fluorescence of CLB01 worms before and during feeding with bacteria expressing <i>C. elegans</i> <i>elt-2</i> dsRNA .....  | 164 |
| Figure 6.7 Graphic analysis of the compounds of interest .....   | 175 |
| Figure 6.8 <i>H. contortus</i> L3 larvae in active and inactive compounds.....   | 177 |
| Figure 6.9 Fluorescence readings of the compounds identified by the <i>H. contortus</i> screen.....  | 179 |
| Figure 6.10 Slower development of CLB01 larvae following exposure to compound.....   | 181 |
| Figure 6.11 GFP fluorescence in CLB01 embryos .....  | 183 |
| Figure 6.12 <i>C. elegans</i> BIS1 ( <i>vit-2::gfp</i> ) worms indicating embryonic GFP fluorescence.....  | 185 |
| Figure 7.1 Phylogenetic tree showing the relationship of CBL-like proteases in <i>H. contortus</i> (Hc), <i>N. americanus</i> (Na), <i>A. caninum</i> (Ac) and <i>C. elegans</i> (Ce)..... | 196 |

## List of Tables

|   |     |
|---|-----|
| Table 1.1 Major anthelmintic classes used in the control of <i>H. contortus</i> .....   | 22  |
| Table 3.1 DNA comparison of the first intron of the BAC protease genes .....  | 59  |
| Table 3.2 Sequence LOGOS and conserved motifs within intronic regions of the BAC genes.....   | 62  |
| Table 3.3 EST data for the BAC genes obtained from Nembase 4.....   | 65  |
| Table 3.4 Percentage identity/similarity at the amino acid level of proteases encoded by <i>H. contortus</i> BAC 18f22 and the AC protease family ..... | 68  |
| Table 3.5 Conserved regions within the BAC and AC CBL cysteine protease families .....  | 74  |
| Table 3.6 Re-naming the genes present on <i>H. contortus</i> BAC18f22.....  | 77  |
| Table 4.1 Percentage identity at the amino acid level of proteases encoded by supercontig_0059492 and the <i>H. contortus</i> HmCP family .....         | 88  |
| Table 4.2 EST data for the genes present on supercontig_0059492 .....   | 89  |
| Table 4.3 Conserved regions within the <i>H. contortus</i> cysteine proteases and the proteases present on supercontig_0059492 .....                    | 93  |
| Table 4.4 Amino acid identities of GCP-7 and the proteases encoded by supercontig_0041161 and supercontig_0059702 .....                                 | 99  |
| Table 4.5 EST data for the genes present on supercontig_0041161 and supercontig_0059702.....  | 101 |
| Table 4.6 Conserved regions within <i>H. contortus</i> GCP-7 and the proteases present on supercontig_0041161 and supercontig_0059702 .....             | 105 |
| Table 4.7 Percentage identity at the amino acid level of CPR-6 in a number of parasitic species .....   | 112 |
| Table 4.8 Conserved regions within the CPR-6 protease in a number of nematodes.....   | 116 |
| Table 5.1 Number of conserved motifs in the 1 kb promoter region of a number of <i>H. contortus</i> genes.....  | 132 |
| Table 6.1 Analysis of the fluorescence levels from the preliminary 4 day test screen .....  | 158 |
| Table 6.2 Analysis of the 22 compounds used in the first compound screen ....   | 166 |
| Table 6.3 Number of progeny in compound and control wells after three days  | 180 |

## Acknowledgements

First I would like to thank my supervisor Dr. Collette Britton for her endless guidance throughout this project and for proof reading this thesis. I also appreciate her support in allowing me time off to play rugby.

Thanks to my assessor Professor Tony Page for his valuable suggestions during our meetings. Thanks also to Dr. Brett Roberts for all her help and advice throughout this project and for making lab days a little less tedious. I would also like to thank the other members of the parasitology group, Professor Eileen Devaney, Alan, Gill, Vicky, and Kirsty for their advice and guidance. For all their help with *H. contortus* genome work I would like to thank Dr. Roz Laing, Professor John Gilleard (University of Calgary, Canada) and Dr. James Cotton (Wellcome Trust Sanger Institute), and for all his help with bioinformatic work, a thank you to Dr. William Weir.

Thanks also to Dr. Debra Woods at Pfizer for making me feel very welcome and for the advice from herself and everyone else in the Veterinary Medicine Research & Development Department at Pfizer, Kalamazoo.

I would like to thank the funding bodies that made this project possible; BBSRC, Biosciences KTN, Pfizer and the University of Glasgow.

A special thank you to my parents, sister, other relatives and friends for their support and useful advice. Also to Dr. Zeeshan Durrani who never left, providing me with company in the office right until the end. Lastly I would like to thank Louise for all her patience with my stresses throughout this PhD and her complete support with this job.

## List of Abbreviations and Symbols

|                   |                                     |
|-------------------|-------------------------------------|
| $\alpha$          | alpha                               |
| AAD               | Amino-acetonitrile derivative       |
| BAC               | Bacterial artificial chromosome     |
| bp                | base pair                           |
| BLAST             | basic local alignment search tool   |
| $\beta$           | beta                                |
| Ce                | <i>Caenorhabditis elegans</i>       |
| CO <sub>2</sub>   | carbon dioxide                      |
| CBL               | cathepsin B-like                    |
| cm                | centimetre                          |
| contig            | Contiguous sequence                 |
| cDNA              | complementary deoxyribonucleic acid |
| °C                | degrees celsius                     |
| DNA               | Deoxyribonucleic acid               |
| DNase             | Deoxyribonuclease                   |
| dNTP              | Deoxyribonucleotide triphosphate    |
| dsRNA             | Double-stranded ribonucleic acid    |
| DMSO              | Dimethyl sulfoxide                  |
| dH <sub>2</sub> O | distilled water                     |
| ELISA             | Enzyme linked immunosorbent assay   |
| EDTA              | Ethylenediaminetetra-acetic acid    |
| EST               | Expressed sequence tag              |
| g                 | gram                                |
| GFP               | green fluorescent protein           |
| Hc                | <i>Haemonchus contortus</i>         |
| h                 | hour                                |
| IME               | intron-mediated enhancement         |
| kb                | kilobase (1,000 bp)                 |
| kDa               | kilodalton                          |
| l                 | litre                               |
| Mb                | megabase                            |
| mRNA              | messenger ribonucleic acid          |
| $\mu$             | micro                               |

|          |  |
|----------|--|
| µg       | microgram  |
| µl       | microlitre   |
| µM       | micromolar   |
| miRNA    | micro ribonucleic acid                                     |
| mg       | milligram  |
| ml       | millilitre   |
| mm       | millimetre   |
| mM       | millimolar   |
| min      | minute   |
| M        | molar  |
| ng       | nanogram   |
| p        | p value; statistical significance                          |
| PBS      | Phosphate buffered saline                                  |
| PAGE     | Polyacrylamide gel electrophoresis                         |
| PCR      | Polymerase chain reaction                                  |
| RACE     | Rapid Amplification of cDNA Ends                           |
| RT       | Reverse transcriptase                                      |
| RT-PCR   | Reverse transcription polymerase chain reaction            |
| RNase    | Ribonuclease   |
| RNA      | Ribonucleic acid   |
| RNAi     | Ribonucleic acid interference                              |
| sec      | second   |
| NaCl     | sodium chloride  |
| SDS      | sodium dodecyl sulphate                                    |
| SDS-PAGE | sodium dodecyl sulphate polyacrylamide gel electrophoresis |
| TF       | transcription factor                                       |
| Taq      | Thermus aquaticus polymerase                               |
| TAE      | Tris-Acetate EDTA  |
| UV       | Ultraviolet  |
| V        | volt   |
| YAC      | Yeast artificial chromosome                                |

# **Chapter 1**

## Introduction



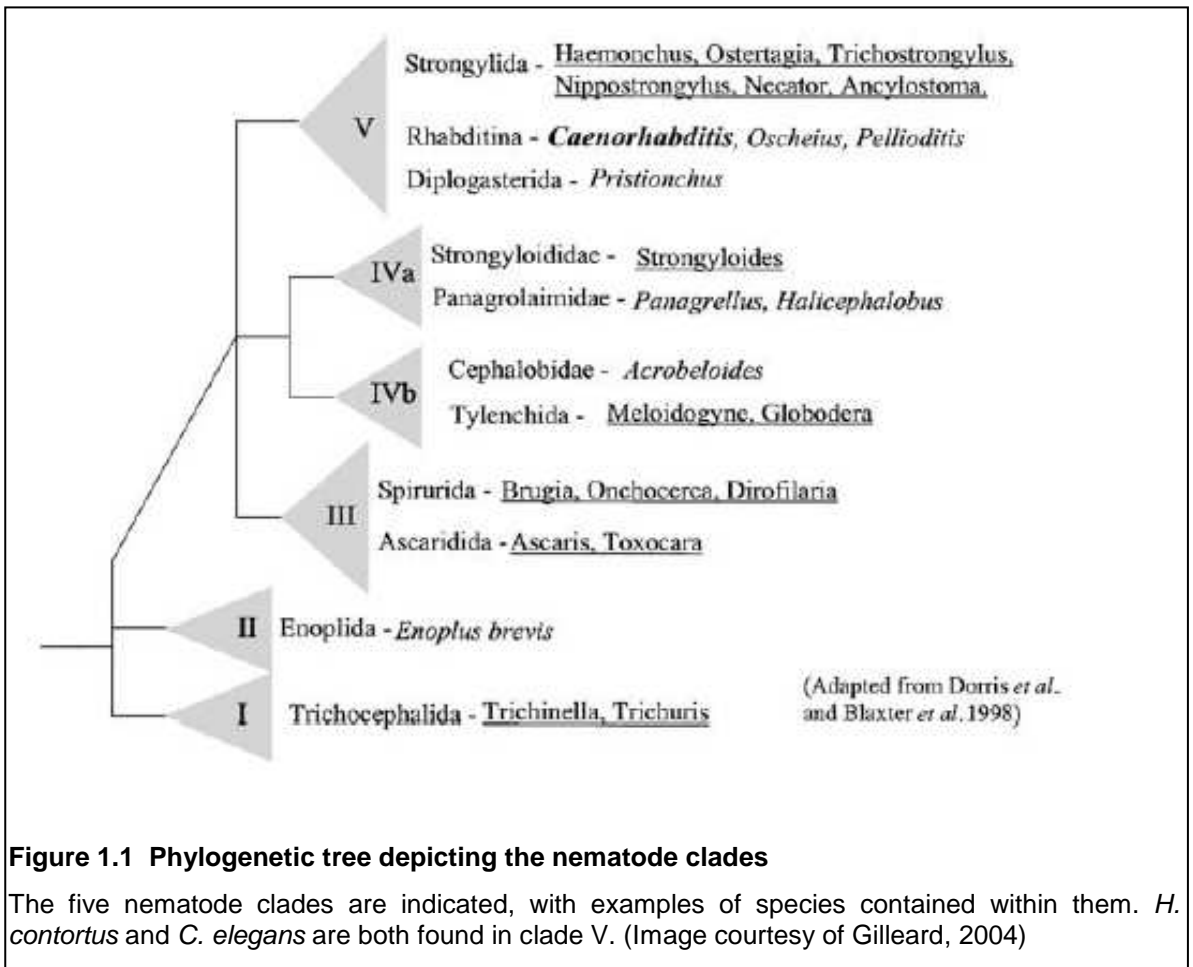
## 1.1 Parasitic nematode infections

### 1.1.1 *Haemonchus contortus*

Many helminths are responsible for parasitic infection in plants, animals and humans. It is estimated that over one billion people are infected worldwide with the majority of these individuals being in the developing countries of the world (Hotez *et al.*, 2008). In the UK alone, endemic parasite infections cost the livestock industry millions of pounds every year (<http://www.knowledgescotland.org/briefings.php?id=75>).

*Haemonchus contortus* is a Strongylid parasite and is one of the most pathogenic nematodes in sheep and goats. It is a blood sucking abomasal parasite that can cause anaemia, weakness, oedema, loss of appetite, diarrhoea and subsequently death. Economically, *H. contortus* is one of the most important parasites of ruminants worldwide, with weight loss resulting in decreased production and in many cases mortality (Githigia *et al.*, 2001). Development of the *H. contortus* free-living stage is enhanced by the presence of warm and wet climatic conditions, indicating why this parasite was traditionally found in more tropical climates (Waller *et al.*, 2006). *H. contortus* is becoming more widespread, an occurrence that may be due to changes in sheep management and the farming industry, climate change and from the overuse/incorrect use of anthelmintic drugs (Coles, 2002). Primary infection comes from the ingestion of infective L3 larvae on the pasture, therefore an increase in the intensity of sheep farming resulting in higher stocking densities can lead to an increase in larval numbers consumed from the pasture (Thamsborg *et al.*, 1996). Global warming resulting in warmer, wetter conditions favouring parasite survival may be a contributing factor to the increase in parasitaemia. As such conditions are favourable for larval development, this can result in a rapid increase in infective larvae on pastures (Kenyon *et al.*, 2009). Compounding the increased exposure of livestock to parasitic nematodes is the problem of anthelmintic resistance. Surveys suggest that 80% of UK sheep farms have resistance to at least one class of anthelmintic drug (Taylor, 2009).

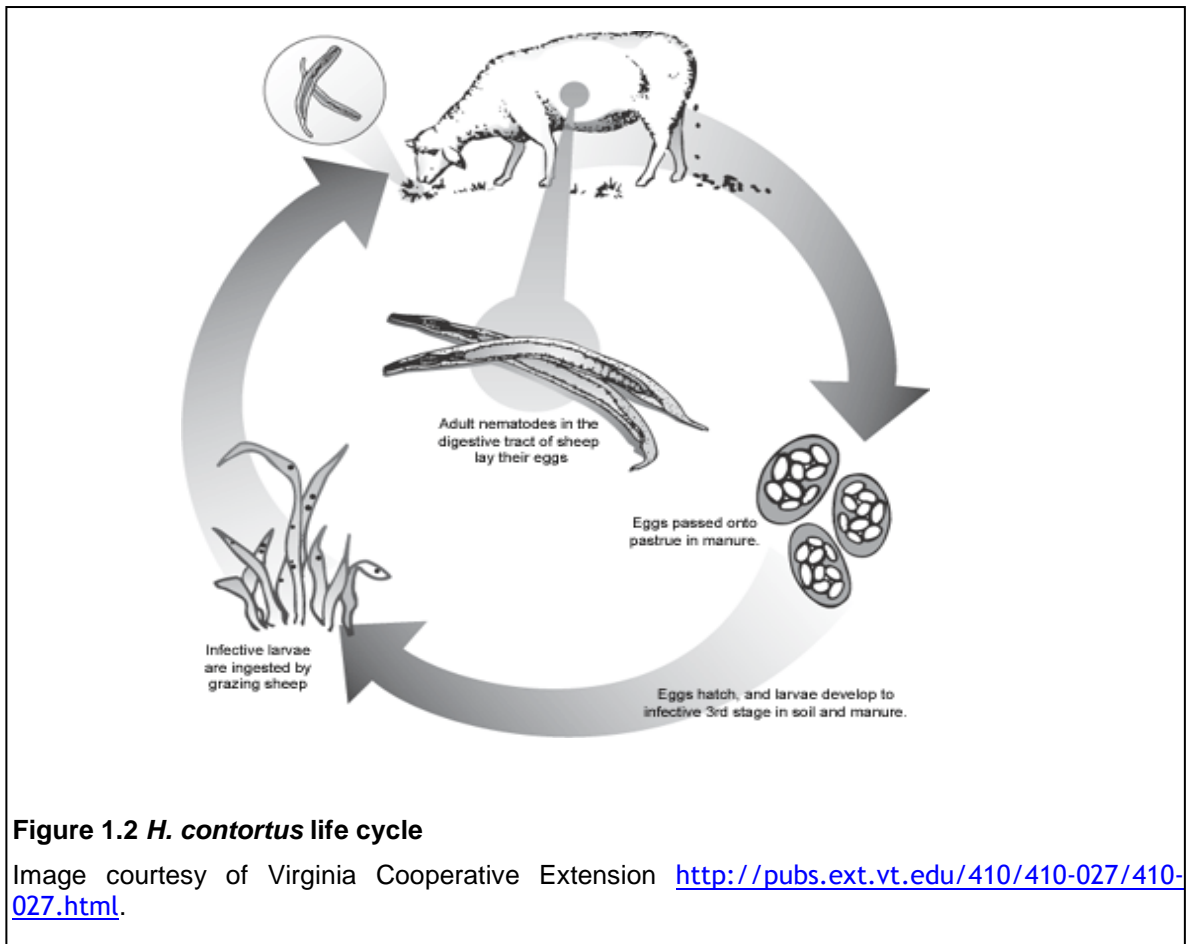
Nematode species can be categorised into clades (Blaxter, 1998) based on small subunit ribosomal RNA gene sequence and other genes. Each clade contains a mixture of free-living and parasitic species except for clade III which contains only animal parasites (Figure 1.1). This phylogenetic analysis has helped group the free-living nematode *Caenorhabditis elegans* in the same clade as *H. contortus*. This indicates that the two species are closely related and suggests *C. elegans* is suitable for the study of *H. contortus* (Gilleard, 2004).



### 1.1.2 *H. contortus* life cycle

Knowledge of the *H. contortus* life cycle (Figure 1.2) is important for the understanding of possible control methods and the correct use of anthelmintic drugs. *H. contortus* has a direct life cycle, and is an obligate parasite, with sheep and goats being the definitive host (Angulo-Cubillan *et al.*, 2010). Initially eggs are passed in faeces onto the pasture. These eggs hatch on the pasture to the L1 stage. Moulting occurs, producing the L2 stage, after the second moult the cuticle is not shed resulting in ensheathed L3 larvae that cannot feed. L3 move from the faeces onto the pasture and are ingested by the host, with the

host required for the completion of the life cycle. After ingestion, L3 larvae exsheath and moult, resulting in L4s in the abomasum. These larvae latch on to the stomach lining and begin to feed, causing bleeding and clot formation around the individual larva. Each worm results in the loss of about 0.05 ml of blood per day through ingestion and from lesions. Thus a sheep with a burden of 5,000 worms could lose up to 250 ml per day, an indication of why anaemia is one of the symptoms of infection (Abbott *et al.*, 2009). After approximately 3 days the mature L4 larvae emerges from the clot and moults to the final adult stage. It is at this point that the adult worm attaches to the mucosa of the abomasum, mates and begins producing eggs. When mature, worms can be up to 2 cm in length and a mature female can produce thousands of eggs per day. *H. contortus* eggs can first be detected in the faeces approximately 14 days after initial larval ingestion (these time periods are important for the use of strategic anthelmintic dosing). Larvae can also undergo arrested development, thus they can survive *in vivo* through periods where pasture survival would not be possible (Waller *et al.*, 2004). Bad pasture management and the pattern of sheep grazing (for example, exposing clean pasture to resistant worms and/or not rotating pasture) (<http://www.fwi.co.uk/academy/article/126167/worms-in-sheep.html>) can lead to a large ingestion of infective larvae. At certain times of the year (e.g. spring) many larvae restart their development at the same time, resulting in a noticeable detrimental effect on the health of the flock (Whittier *et al.*, 2009).



### 1.1.3 The use of anthelmintic drugs to treat parasitic infection

The predominant method of control for *H. contortus* is the use of broad spectrum anthelmintic drugs of which there are three major classes; Benzimidazoles (BZ, group 1), Imidazothiazoles and Tetrahydropyrimidines (group 2) and Avermectins and Milbemycins (group 3). Drugs contained within families have the same mode of action, with many of these drugs targeting the parasite nervous system (Sleigh, 2010). Anthelmintic drugs are generally administered by oral drench however other methods such as slow release bolus, injection or in-feed methods are also available. There are both advantages and disadvantages to each delivery method (Taylor, 1999).

Group 1 anthelmintics inhibit microtubule formation by binding to the tubulin precursors and preventing polymerisation through a process known as 'capping' (Oxberry *et al.*, 2001). Microtubules are intracellular organelles composed of  $\alpha$ -tubulin and  $\beta$ -tubulin proteins. They have a wide range of functions including formation of the mitotic spindle, maintenance of cell shape and intracellular transport, thus are essential for normal growth and development. Tubulin

contains three binding domains and it is the colchine binding domain at which benzimidazoles bind, forming a tubulin-benzimidazole complex (Lacey, 1988). Evidence suggests that benzimidazoles bind to the  $\beta$ -tubulin gene, as mutations in this sequence confer resistance (Drogemuller *et al.*, 2004). Other proposed ideas include benzimidazoles binding to the  $\alpha$ -subunit but still being regulated by  $\beta$ -tubulin (Banerjee and Luduena, 1992).

Group 2 anthelmintics are nicotinic agonists which target the nicotinic acetylcholine receptors (nAChRs) at nematode neuromuscular junctions. These anthelmintics have more potent activity than acetylcholine at the receptors and cause the opening of ligand-gated receptors on the muscle cell. Anthelmintic drugs are not rapidly degraded resulting in a prolonged channel opening and continued  $\text{Ca}^{2+}$  entry (Sleigh, 2010). The resulting contraction of the muscle fibres causes a rigid paralysis of the worm (Harrow and Gratton, 1985).

Group 3 anthelmintics are a group of Macrocyclic lactones that are produced from the fermentation of *Streptomyces* species (Taylor, 1999). This group is able to target the GABA receptors in both vertebrates and invertebrates, and the glutamate-gated chloride channels in invertebrates. Macrocyclic lactones target these ligand-gated ion receptors which function in neuronal and muscular systems, and paralyse the parasite by increasing permeability of the muscle to  $\text{Cl}^-$  (Cully *et al.*, 1996).

It has been over 25 years since the last anthelmintic was released on to the market. Within this period two classes, cyclodepsipeptides and paraherquamides have emerged but as yet neither have produced a marketable product (Ducray *et al.*, 2008). More recently however a new class, the amino-acetonitrile derivatives (AADs), has been shown to provide anthelmintic activity (Kaminsky *et al.*, 2008a; Kaminsky *et al.*, 2008b). This is a synthetic class of drug that has provided activity against nematodes. Work has been carried out in *C. elegans* to determine the mode of action of the AADs and an *acr-23* gene was identified as a major target. In *H. contortus* the *monepantel-1* (*Hco-mptl-1*) gene is thought to be a potential target. Both these genes are part of the ligand-gated ion channel superfamily and are nAChR subunits (Rufener *et al.*, 2010). These drugs work by a novel mode of action and are effective against those isolates already resistant to conventional drug classes, a necessity for any new anthelmintic. The main

drug used in studies is Monepantel (AAD 1566) as it targets resistant isolates, has a high efficacy and is well tolerated *in vivo*. It causes contraction of the body wall muscles leading to parasite paralysis and death (Kaminsky *et al.*, 2009). This drug has been developed by Novartis Animal Health and was released under the trade name Zolvix in 2010. In addition to the AADs was the discovery of the spiroindoles (SI) another new class (Lee *et al.*, 2002). The SIs work as cholinergic neuromuscular blockers to cause flaccid paralysis and as the target binding sites are different from other classes it is effective against worms resistant to other classes (Little *et al.*, 2010). 2-deoxy-paraquamide (derquantel) was the first studied and when used solely was a mid-spectrum anthelmintic, for this reason it was combined with abamectin and trialled in field (Little *et al.*, 2011). STARTECT (Pfizer Animal Health) was subsequently released into the market in 2011. As these are the first new anthelmintics in 25 years, care must be taken to preserve their efficacy for as long as possible as there is now widespread resistance to the other three classes of anthelmintic drug.

#### **1.1.4 Anthelmintic resistance**

Drug resistance is defined by the change in the pharmacodynamics of a drug (Lacey, 1990). There is now nematode resistance to the three main classes of anthelmintic currently available. The resulting decrease in production and increase in mortality is an increasing economic threat to the livestock industry, thus alternative control approaches are required. Development of the new drugs Monepantel and STARTECT are two attempts to combat this problem. There are many factors contributing to anthelmintic resistance, examples include; overuse, incorrect dosage and unnecessary dosing (Prichard, 1990). Once resistance has developed to a compound within a group, the efficacy of other members within the group is also decreased due to the conservation of mode of action between members within a group.

*H. contortus* was one of the first parasites to demonstrate anthelmintic resistance in the early 1960s (Conway, 1964), to the drug Phenothiazine a member of the Benzimidazole family. Thiabendazole was then introduced to the market, but subsequent overuse resulted in its inefficacy within just a few years. Widespread resistance to the Benzimidazole family had become apparent

by the 1970s. This pattern of resistance was also followed by the newer anthelmintic classes; Imidazothiazoles and Tetrahydropyrimidines and Avermectins and Milbemyicins (Table 1.1). By the early 1980s there were reports of multiple drug resistant (MDR) worms demonstrating resistance to drugs in different classes. Of these, resistance to the Benzimidazole class of drug is the most understood (Kaplan, 2004). There are two isotypes of  $\beta$ -tubulin that have been identified in parasitic nematode species. Using *C. elegans* Kwa *et al.*, (1995) showed that mutation of the  $\beta$ -tubulin isotype-1 gene resulted in resistance to Benzimidazoles and that this could be reversed by introduction of *H. contortus*  $\beta$ -tubulin. Resistance to group 2 anthelmintics can occur in several ways, either by a decrease in number or sequence mutation of the nAChRs in muscle, or by inefficacy after long periods of exposure to the drug, resulting in either complete recovery or partial paralysis (Lewis *et al.*, 1980).

| Drug  | Year of drug approval | First report of resistance published |
|---|-----------------------|--------------------------------------|
| <b>Benzimidazoles</b>                         |                       |                                      |
| Thiabendazole                                 | 1961                  | 1964                                 |
| <b>Imidazothiazoles-Tetrahydropyrimidines</b> |                       |                                      |
| Levamisole                                    | 1970                  | 1979                                 |
| <b>Avermectins-Milbemyicins</b>               |                       |                                      |
| Ivermectin                                    | 1981                  | 1988                                 |
| Moxidectin                                    | 1991                  | 1995                                 |

**Table 1.1 Major anthelmintic classes used in the control of *H. contortus***  
 The period of time between drug approval into the market and resistance being reported in sheep is indicated. (Adapted from Kaplan, 2004)

As yet there is no recorded resistance to Monepantel but, as indicated for the other classes, conversion to resistance is so rapid that money spent on the formation of new drugs may be considered economically unviable. The increasing information on parasite gene sequence, function and regulation can be exploited to understand mechanisms of drug resistance and to develop novel ways to interfere with parasite development and survival.

### 1.1.5 Alternative approaches for *H. contortus* control

In view of the problem of drug resistance, alternative control measures are required. There is only one successful and commercially available vaccine used for the treatment of parasitic nematodes to date, the Bovilis Husk vaccine. This vaccine contains attenuated larvae and is used against the bovine lungworm *Dictyocaulus viviparus* (Sharma *et al.*, 1988).

Development of a potential vaccine against gastrointestinal parasites will require in depth knowledge of the target proteins of interest. The gut of parasitic blood-feeding nematodes contains a number of proteins that are able to confer protective immunity during the blood-feeding process (Jasmer and McGuire, 1991). Work carried out by Munn *et al.*, (1987) showed that young animals immunized with native *H. contortus* gut membrane proteins showed a significant decrease in worm burden after challenge infection.

There are a number of potential protective antigens that are considered possible vaccine targets, including the H11 aminopeptidase family, the H-gal-GP complex (a complex of metallo and aspartate proteases) and the HmCP family (Knox *et al.*, 2003). These antigens are all hidden, meaning they are not recognised by the host even after exposure to infection (Smith *et al.*, 1997). As *H. contortus* is a blood feeder, the surface of the parasite intestine will be exposed to host immunoglobulin and thus to potential antibodies induced by vaccination. Significant effort has been made to target and potentially interfere with their proposed digestive function. Immunity provided by hidden antigens is different to that from natural infection. In natural infection, if an animal is exposed to the parasite for a number of weeks it will subsequently be able to prevent most larvae establishing due to naturally generated antibodies (Schallig *et al.*, 1997). This is in contrast to animals immunised with hidden antigens in which there would be no or little boosting of the immune system by natural infection, thus booster vaccinations would be required. In order for vaccination to be successful, the formulation, cost effectiveness and route of administration are all important considerations (Emery and Wagland, 1991).

Although H11 and H-gal-GP can induce high levels of protection in native form, attempts to vaccinate sheep with recombinant proteins have proved far less



successful (Cachate *et al.*, 2010;Munn *et al.*, 1997). This is thought to be due to differences in conformation and/or glycosylation and has led to work using *C. elegans* as a suitable expression system for *H. contortus* and *T. circumcincta* vaccine candidates (Murray *et al.*, 2007) (Roberts and Britton, unpublished data; Longhi, Nisbet, Britton and McNeilly, unpublished data). In the meantime, attention has re-focussed on vaccination with native *H. contortus* gut antigens, through improvements to the extraction process. While this still requires a supply of parasites for antigen preparation, recent field trials have proved very successful (Smith *et al.*, 2000;Smith *et al.*, 2003).

Screening using antibody to a 'protein enriched' gut membrane extract resulted in the identification of three cDNAs encoding cathepsin B cysteine proteases, HmCP1, 4 and 6. These HmCP genes are developmentally regulated and are only expressed in blood feeding stages (Skuce *et al.*, 1999). A number of cysteine protease genes were previously identified in a US strain of *H. contortus*, namely the AC protease gene family (Pratt *et al.*, 1990) and *gcp-7* (Rehman and Jasmer, 1999). Work has shown that the HmCP genes are distinct from those previously isolated, as at best they only share a 60% amino acid identity (Skuce *et al.*, 1999).

Protective antigens have also been isolated from excretory-secretory (ES) components obtained through *in vitro* incubation of worms. These antigens are considered to be excreted from the parasite and thus available to the host mucosal lymphoid cells and lymph nodes via the circulatory system. These naturally occurring antigens elicit both a cellular and antibody response (Bakker *et al.*, 2004). Schallig *et al.*, (1997) examined two ES proteins of 15 and 24 kDa from *H. contortus* adult worms and subsequent analysis indicated that a strong antibody response was generated against these antigens, with over 70% protection rates recorded.

The identification of a single protein that provides protection would be ideal for commercial vaccine development. However this may be very difficult to achieve, thus looking for a combination of proteins that together may provide significant protection and/or protection against a number of parasites would also be of great benefit (Munn *et al.*, 1997). For commercial vaccine production, development of a suitable recombinant protein expression system is a priority.

### 1.1.6 Nematode genome sequencing

Sequencing of the *C. elegans* genome has sparked great interest in the study of the human genome and provided better tools for genome analysis (<http://www.hgc.jp/english/software.html>). This has led to the initiation of genome sequencing projects for a wide range of other smaller organisms (<http://www.sanger.ac.uk/>). The free-living nematode *C. elegans* was the first multi-cellular organism to have its genome sequenced and this data has been widely used in comparative studies (Sulston *et al.*, 1992). Additionally there was a wealth of genetic data available indicating that *C. elegans* could be used as a model system for studying development, neurobiology, signalling and behaviour in other multicellular organisms (Brenner, 1974).

Sequencing work began with the creation of a physical map of the *C. elegans* genome. This was only possible through communication of a number of labs working with the *C. elegans* nematode. The aim was to have access to any segment of the genome and provide a start point for large scale sequencing (Coulson *et al.*, 1986). After assembly of the sequenced cosmids, gaps were still present, largely due to the size constraints in the cloning system, with a 50 kb fragment the maximum length of an individually sequenced region. It was not until the development of yeast artificial chromosome (YAC) clones that the majority of the gaps could be filled. Unlike original cloning vectors, YACs can contain fragments of up to several hundred kb (Burke *et al.*, 1987). The *C. elegans* genome was completed in 1998 (The *C. elegans* Sequencing Consortium, 1998), is ~97 Mb in size and encodes over 19,000 genes. The importance of this data not only to *C. elegans* but to other organisms is indicated by the fact that more than 40% of the predicted protein products were found to be homologous to those identified in other organisms. Therefore, information from *C. elegans* genetic mutants and gene knockouts can provide a first indication of the potential role of conserved genes.

The *C. briggsae* genome sequence is also available (98% complete) allowing a direct comparison between the two *Caenorhabditis* species. These nematodes diverged approximately 100 million years ago, but have similar genome sizes. Gene arrangement on chromosomes is also well conserved with 96% of genes

showing synteny (Stein *et al.*, 2003). The information provided by the *C. briggsae* genome has enabled improved gene annotation in *C. elegans*. In addition, analysis of sequence conservation and divergence has helped identify functional domains within coding and non-coding regions, such as motifs in promoter and 3' UTR sequences.

A draft genome assembly and annotation was published for the filarial nematode *Brugia malayi* in 2007. It is an estimated 90-95 Mb in size, similar to that of *C. elegans* and *C. briggsae* and was obtained through whole-genome shotgun (WGS) sequencing (Ghedin *et al.*, 2007), involving the piecing together of overlapping reads to give one continuous sequence. WGS is used for larger fragments by sequencing from opposite ends on different strands. It was identified that there is very little synteny between *B. malayi* and *C. elegans* and that *B. malayi* has a far fewer number of estimated genes than those found in *C. elegans*, however sequencing is still only 80% complete. Having access to a number of nematode genomes enables large-scale comparative studies which, in time, will provide information about nematode developmental pathways and molecules that can be used as possible control targets (Ghedin *et al.*, 2007).

The genome of *H. contortus* is currently being sequenced. This work is being carried out at the Wellcome Trust Sanger Institute and is just one of the many helminth sequencing projects currently underway ([http://www.sanger.ac.uk/cgi-bin/blast/submitblast/h\\_contortus](http://www.sanger.ac.uk/cgi-bin/blast/submitblast/h_contortus)). Current work suggests the *H. contortus* genome may be over four times the size of original estimates (200-300 Mb) (Gilleard lab, unpublished work). The initial estimate of 50 Mb was based on flow cytometry (Leroy *et al.*, 2003) and is thought to be an underestimate. As mentioned previously, both *H. contortus* and *C. elegans* are clade V nematodes and phylogenetic analysis suggests they diverged approximately 400 million years ago (Laing *et al.*, 2011). Thus there are likely to be similarities in gene sequence. The complete sequencing of the *H. contortus* genome would enable all genes to be identified. As this would be the first clade V parasitic nematode for which complete genome data is available it would allow a full genomic comparison between *H. contortus* and *Caenorhabditis* species.

### 1.1.7 *Caenorhabditis elegans* as a model for parasitic nematodes

Parasitic nematodes such as *H. contortus* are not ideal for *in vitro* studies; they have complex lifestyles and are difficult to culture and maintain. It is for this reason that research on *H. contortus* and other parasitic nematodes has made use of the free-living nematode *C. elegans* (Bürglin *et al.*, 1998). *C. elegans* is a soil living nematode and its growth requirements are relatively basic; warm humid temperature, oxygen and bacteria as food. The *C. elegans* life-cycle is very rapid and takes just 3.5 days at 20°C with each worm producing up to 300 progeny in just 4 days, thus providing a very useful resource for large-scale studies (Byerly *et al.*, 1976). Sequence as well as functional information on *C. elegans* genes (<http://www.wormbase.org>) also provides an important resource for studies on parasitic nematodes, particularly for clade V nematodes such as *H. contortus* and hookworms.

*C. elegans* worms are amenable to transformation making it possible to identify the expression pattern of genes spatially and temporally. More recently the roles of genes within specific tissues have been studied using gene-specific promoters to drive expression, for example *Hc-cpr-1* (Britton *et al.*, 1998) *Hc-elt-2* (Couthier *et al.*, 2004). These types of studies have been possible due to work carried out by Fire *et al.*, (1990) which saw the creation of a number of vectors, containing various promoters and marker genes and which have been improved over time. Parasitic gene promoter regions and functional genes can also be introduced into these vectors and then into *C. elegans* worms creating transgenic lines that can be easily cultured and studied *in vitro*. This has helped demonstrate the expression pattern of genes and show that similar gene regulatory mechanisms exist in *C. elegans* and parasitic nematodes (Britton *et al.*, 1999). In addition, rescue of *C. elegans* mutant phenotypes can be used to demonstrate conserved gene function (Britton and Murray, 2002; Gillan *et al.*, 2009).

Analysis of gene function in *C. elegans* and in other organisms has been revolutionised by the development of RNA interference (RNAi). This is a mechanism used to silence genes in a sequence specific manner and is achieved by introducing dsRNA into the worm (Fire *et al.*, 1998). Work has been

successfully carried out using *C. elegans* due to its ease of culture *in vitro*. In *C. elegans* RNAi is effective not only when the dsRNA is injected but also by soaking the worms (Tabara *et al.*, 1998) and by feeding the worms on *Escherichia coli* expressing dsRNA (Fire *et al.*, 1998). However *C. elegans* seems to be an exception in the ease with which genes can be silenced. In *C. briggsae*, RNAi can be induced only by injection of dsRNA, a difference that is speculated to be due to different SID-2 transporter proteins (Winston *et al.*, 2007). In parasitic nematodes *in vitro* culture alone is difficult, and thus RNAi approaches which effectively deliver dsRNA to worms and allow analysis of any resulting phenotype have proved difficult (Geldhof *et al.*, 2007). Soaking is the main method used, but achieving sufficient uptake has proved difficult. Alternative approaches such as microinjection of the small larval stages is not easy due to poor survival, and electroporation did not improve gene knockdown (Geldhof *et al.*, 2006). Differences in transport may also explain RNAi deficiencies in parasitic nematodes (Winston *et al.*, 2007). Future work such as the introduction of the *sid-2* gene or other *C. elegans* transporters into parasitic nematodes may lead to better RNAi effects. Transgenic methods have been developed for *Strongyloides* species and *B. malayi* thus making such studies potentially feasible (Li *et al.*, 2006; Xu *et al.*, 2011).

## 1.2 Gene regulation

### 1.2.1 Gene regulation in nematodes

Gene regulation in nematodes can occur via many mechanisms. Currently the most well-studied include direct regulation by transcription factors (TFs) and microRNAs (miRNAs) (Carthew, 2006; Hobert, 2008). TFs are proteins that bind to specific regions of DNA, termed cis-regulatory DNA elements, and cause either a positive or negative effect on gene transcription. They contain at least one DNA binding domain which targets the upstream region of the genes to which they bind (Reece-Hoyes *et al.*, 2005). Identification of gene regulatory elements has traditionally relied on the “promoter bashing” approach. This involved deletion and mutation of regions thought to be important for regulation. The sequence information that is now available for a number of *Caenorhabditis* species has enabled the prediction of putative protein coding genes in *C. elegans*, and has

made identification of conserved regulatory elements within the promoter regions of these genes possible (Gaudet and McGhee, 2010).

Haerty *et al.*, (2008) carried out an extensive study looking at putative TF activity in *C. elegans*. 998 genes were identified as being under the control of TFs. TF networks have been created to help understand differences in gene expression. These networks identified the relationships between gene targets and their TF proteins, and enabled functionally important TFs to be highlighted (Reece-Hoyes *et al.*, 2005). One TF which has been well studied in *C. elegans* is HSF-1, a heat-shock factor that responds to heat and stress. When HSF-1 is over-expressed there is an increase in lifespan, however if HSF-1 is decreased the opposite is true (Hsu *et al.*, 2003). One major advantage of using *C. elegans* for the study of TFs is the ability to knock-out TFs in the genome to identify their function (Gaudet and McGhee, 2010). Given the importance of TFs in gene regulation, loss of activity often results in phenotypic effects.

In addition to transcriptional regulation by TFs, many genes can also be subject to post-transcriptional regulation by microRNAs (miRNAs). TFs themselves can be regulated by miRNAs, establishing feedback control loops (Martinez *et al.*, 2008). miRNAs are single stranded RNAs that are approximately 22 nucleotides in length and regulate gene expression in protein-coding genes. miRNAs were first identified in *C. elegans* and their mechanism of action demonstrates similarities to the RNAi pathway. Unlike RNAi in which the target mRNA is always degraded, many miRNAs work by repressing the translation of their target RNAs with only some causing the degradation of the target RNAs (Bartel, 2004). The first miRNA identified was *lin-4* which targeted the LIN-14 protein (Lee *et al.*, 1993). In this mechanism of gene regulation, the *lin-4* gene binds to sites present at the 3' UTR of *lin-14*. Subsequent inhibition of the LIN-14 protein occurs, which affects early larval development. *let-7* was the second miRNA gene to be discovered and results in the inhibition of LIN-41 expression which is responsible for late larval development. Subsequently *let-7* was identified in a number of other species and to date hundreds of miRNA genes have been identified in plants and animals (Bagga *et al.*, 2005). More recently, miRNAs have been identified as having a role in drug resistance, with a large focus in tumour cell biology (Zhang *et al.*, 2007). If miRNAs play a part in tumour drug efficacy they may also affect other drugs. One proposal is that the changes in gene expression levels, associated

with miRNAs could have a role in drug resistance in parasitic nematodes (Devaney *et al.*, 2010).

## 1.2.2 Regulation of gut expressed genes in nematodes

The nematode intestine is a remarkable structure and has been shown to be involved in lifespan and ageing (Libina *et al.*, 2003). Libina *et al.*, (2003) identified that activity of the TF DAF-16 in the intestine increases worm lifespan by both positive and negative regulation of gene expression. Thus detailed knowledge of the *C. elegans* intestinal composition is an invaluable resource.

Due to the importance of the *C. elegans* intestine in survival and development, regulation of a number of gut genes has been studied experimentally. Work carried out to date has identified the presence of GATA transcription factors involved in the control of *C. elegans* gut gene promoters. GATA factors are transcriptional activators that bind to GATA sequences in DNA. There are six classes of mammalian GATA factor, named GATA 1-6, all of which contain a distinct pair of zinc finger domains. In *C. elegans* a GATA-1 like gene termed *elt-1*, was initially identified and found to be responsible for control of *C. elegans* gut genes, for example *ges-1* (Kennedy *et al.*, 1993). Probing of *C. elegans* cDNA expression libraries identified a second GATA factor termed *elt-2*. ELT-2 appears to be most closely related to GATA-5 and unlike the other GATA factors only contains one zinc finger (Hawkins and McGhee, 1995). A recent study looking at putative TFs in *C. elegans* identified 9 GATA-type TFs which are highly conserved in other nematode species *C. briggsae* and *Caenorhabditis remanei* (Haerty *et al.*, 2008).

There are three GATA-type TFs expressed in the intestine of the adult *C. elegans* worm; ELT-2, ELT-4 and ELT-7. Experimentally, ELT-2 has been shown to be the major transcription factor (McGhee *et al.*, 2007). Evidence that the ELT-2 GATA TF is essential for gut gene regulation and gut development was determined by studies on the gut specific *ges-1* esterase gene (Fukushige *et al.*, 1998). *elt-2* also plays a role in the activation of endodermal differentiation and specific gene expression. *end-1* and *end-3* (Stainier, 2002) are involved in the initiation phase of gut development and *elt-2* is required for progression of development. Work by McGhee *et al.*, (2007) additionally identified 108 non-GATA TFs that

were expressed in the *C. elegans* adult intestine. RNAi has been applied to determine the roles of these TFs; to date, *elt-2* RNAi has been found to have a significant effect on gene function.

The intestine has been identified as an important source of protective antigens in *H. contortus*, with some of the enzymes found being similar to those found in *C. elegans*. Some of these enzymes are under the control of GATA-type TFs in *C. elegans*, thus using *C. elegans* as a model to study *H. contortus* promoters and TFs can help identify regulatory elements and TFs controlling expression of *H. contortus* gut genes. A better understanding of gut gene regulation in *H. contortus* and other nematodes has important relevance to interfering with gut activity in general and may lead to new approaches to parasite control.

### 1.3 Aims and Objectives

The overall objective of this project was to study *H. contortus* gut gene regulation and function, and to determine whether ELT-2 is a master regulator of gut gene expression in this parasitic nematode.

The main aims of this work were to;

- Identify and characterise potential gut expressed genes from *H. contortus* by annotation of the available genome data.
- Carry out bioinformatic promoter region analysis to identify potential regulatory motifs controlling gut gene expression.
- Identify small molecule compounds that interfere with gut gene expression and potentially target the ELT-2 GATA TF in *C. elegans* worms, using transgenic *C. elegans* as a screening tool (collaboration with Pfizer Veterinary Medicine, Kalamazoo, USA).



## **Chapter 2**

### Materials and Methods

## 2.1 *C. elegans* methods

### 2.1.1 Culture and maintenance

*C. elegans* worm strains used in this thesis were obtained from the *Caenorhabditis* Genetics Centre (CGC) (<http://www.cbs.umn.edu/CGC/strains/>) unless otherwise stated. For optimum development, *C. elegans* worms were maintained on NGM agar plates (Appendix 1) seeded with an *E. coli* OP50 strain at 20°C. In some instances plates became contaminated with either an overgrowth of bacteria or with fungi, bleaching was used to clean cultures. The plate to be cleaned should contain a sufficient number of gravid hermaphrodites. The plate was washed with 1 ml of sterile water and the worm/water mixture transferred to a 1.5 ml eppendorf. 0.5 M NaOH and 1% bleach (Sodium Hypochlorite) were added and the solution vortexed for a few seconds every 2 min for 10 min. The solution was centrifuged at 13,000 g for 30 sec to pellet eggs. The pellet was washed with water and this was repeated until the water appeared clean. An end volume of approximately 100 µl of embryos was transferred onto a seeded OP50 plate and cultured under standard procedures.

*C. elegans* worm strains can be maintained for long periods of time. Once a plate has cleared of bacteria it can be sealed with parafilm and stored at 15°C. At this temperature and with no food the worms become arrested and can be kept in this state for months. Alternatively, worm strains can be frozen and stored at -80°C. Worm plates just cleared of bacteria were washed with equal volumes of M9 buffer (Appendix 1) and freezing solution (Appendix 1). This solution was split between two 1.5 ml screw top eppendorf tubes and placed in a polystyrene box at -80°C overnight to freeze slowly and prevent damage to the worms, and was subsequently transferred to a freezer box for storage. Worms were available for use as required by thawing and transferring the contents onto seeded NGM plates.

### 2.1.2 Transformation of *C. elegans* by microinjection

The microinjection procedure was as described in WormBook ([http://www.wormbook.org/chapters/www\\_transformationmicroinjection/transformationmicroinjection.html](http://www.wormbook.org/chapters/www_transformationmicroinjection/transformationmicroinjection.html)). DNA was injected along with a marker gene into the gonads of adult hermaphrodites. pRF4 (*rol-6*) (Kramer *et al.*, 1990) and p76-16B (*unc-rescue*) (Bloom and Horvitz, 1997) marker gene plasmids were used in this study. Introduction of *rol-6* to wild-type worms has a visible rolling effect and *unc-76* rescue was achieved by injecting the rescue plasmid into the DR96 strain.

DNA for injection was purified using the QIAGEN Plasmid mini kit. The injection mix contained 10-50 ng/ $\mu$ l of DNA (concentrations dependent on plasmid) and 100 ng/ $\mu$ l of the marker gene in distilled water. The mix was centrifuged at 13,000 g for 15 min and the supernatant transferred to a clean tube to eliminate any substrates that may block the needles.

Needles were prepared from borosilicate glass capillaries (1.2 mm diameter, Harvard Apparatus) using a 773 APP Micropipette Puller (Campden Instruments Ltd). They can either be loaded using a standard bench pipette or a mouth pipette and a volume of approximately 1  $\mu$ l is sufficient for injection of tens of worms. Needles were placed into a needle holder attached to a Zeiss Axiovert S100 Inverted Differential Interference Contrast (DIC) microscope and a pressure of 40 psi applied to the needle using a foot pedal.

For injection, healthy young adult worms were placed onto a agarose injection pad containing a drop of mineral oil, to prevent dehydration, and then positioned so that the gonad was visible. Agarose injection pads were made by pipetting a small volume of warm 2% UltraPure Agarose (Invitrogen) solution onto a glass cover slip and then using another to flatten it, giving a flat circular pad. Pads were then baked at 40°C for 2 hours before use. Once the needle tip was inserted into the worm, filling of the gonad was observed when DNA solution was injected. Worms were placed onto OP50 seeded NGM agar plates, with a small volume of M9 for recovery. After 3 days the F1 generation were of an age at which phenotypic markers could be observed. F1 transformants were

transferred onto fresh OP50 seeded NGM agar plates and allowed to egg lay. Stable F2 transformants were then selected.

### 2.1.3 Staining and viewing *C. elegans* worms

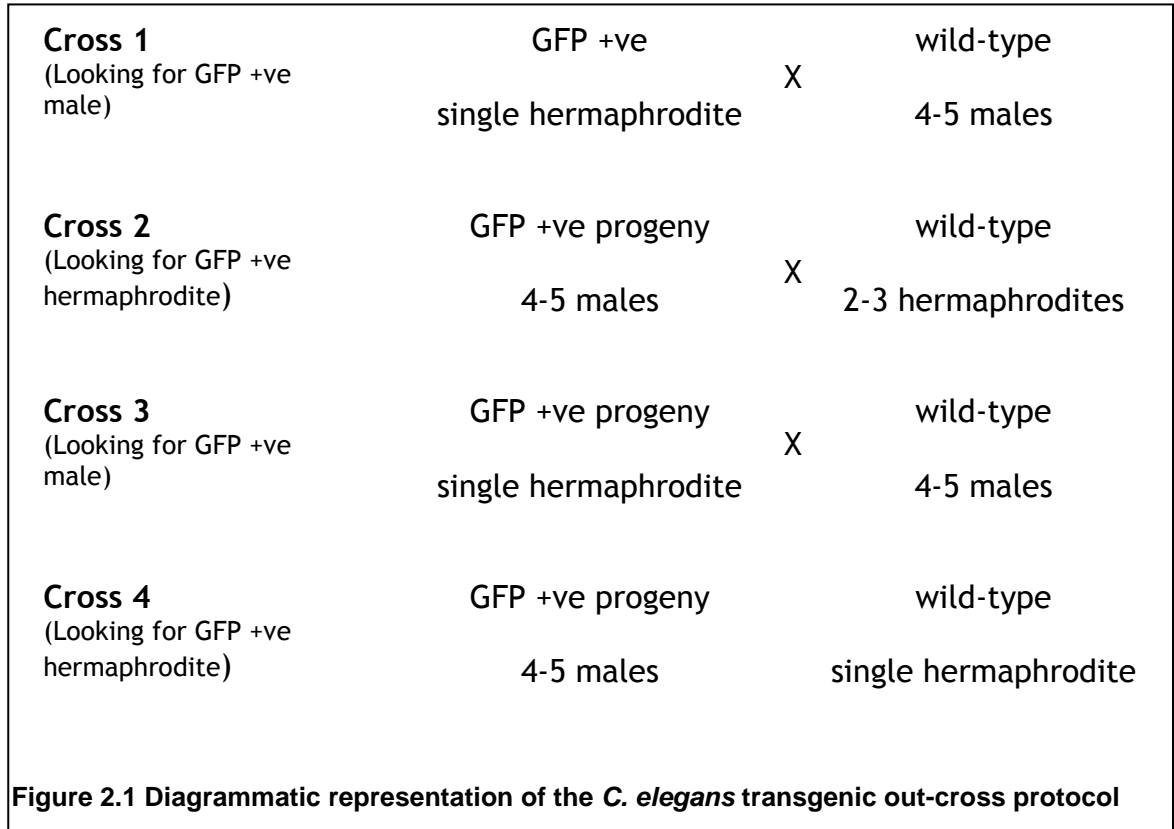
In this study, work carried out with transgenic lines had either Green Fluorescent Protein (GFP) or Lac-Z as a marker. GFP can be viewed using a bench microscope with a UV light and requires no staining for visualisation. Conversely, Lac-Z is only visible after staining with an X-gal stain (Appendix 1) for  $\beta$ -galactosidase activity. Worms for Lac-Z staining were washed off plates using an M9/0.001% Triton solution (Appendix 1) into a 1.5 ml eppendorf. The tubes were centrifuged at 13,000 g for 1 min to pellet worms and the excess supernatant removed leaving approximately 100  $\mu$ l. 1.25% gluteraldehyde (Sigma Aldrich) was added and left for 15 min. The worms were subsequently washed in the M9/Triton solution three times to remove all gluteraldehyde. An approximate volume of 50  $\mu$ l was left after the last wash and transferred onto a glass slide. Slides were left to air dry and then placed in acetone at  $-20^{\circ}\text{C}$  for 3-5 min for fixation of worms. Once dry, approximately 100-200  $\mu$ l of X-gal stain was added to the slides and left to develop in a humid chamber for a minimum of 1 hour, with overnight incubations giving stronger results. Worms were examined using an Axioscop 2 plus microscope (Zeiss) under bright field light.

### 2.1.4 Transgene integration

The method used is based on John Kim's adaptation of the Greenwald and Hobert Labs' integration protocol ([http://www.elegansfusion.org/index.php?option=com\\_content&task=view&id=22&Itemid=43](http://www.elegansfusion.org/index.php?option=com_content&task=view&id=22&Itemid=43)). Integration is used to achieve 100% transmission of a transgenic line with an original transmission rate of <100%. The *Ce-cpl-1::gfp+unc-76* rescued transgenic line used in this study had a transmission rate of ~70% and was selected for integration based on its potential use in *elt-2* RNAi studies. 5 transgenic L4 hermaphrodites were picked onto an OP50 seeded NGM agar plate, and this procedure repeated to give 30 plates in total. The plates were placed in a Stratagene UV crosslinker (Stratalinker) with the lids removed and irradiated using 350  $\mu$ J x100. The plates were left to starve for approximately 10-14 days at

20°C. Each plate was numbered 1-30 and from each starved plate 4 small areas were removed and placed onto fresh OP50 seeded NGM agar plates. These plates were numbered according to the plate from which they came, e.g. 1.1, 1.2, 1.3, 1.4, giving a total of 120 plates. These plates were left for 1-2 days and 5 transgenic hermaphrodites were picked from each plate. The adults were put onto smaller 3 cm seeded plates and left to egg lay. Approximately 60 eggs per plate is ideal and the adults were subsequently removed. Each plate was given a letter and a number corresponding to the plate of origin, e.g. 1.1a, 1.1b, 1.1c, 1.1d, 1.1e, giving a total of 600 small plates containing eggs. These plates were left for a few days and then observed and any plates with 100% transgenic animals are fully integrated.

Because irradiation results in chromosome disruption, mutations can arise and for this reason integrated lines were out-crossed by mating with wild-type males. *C. elegans* males occur at a very low percentage within a population. Male wild-type worms were picked onto seeded NGM agar plates along with a number of young hermaphrodites to obtain a supply of males for use. The out-cross protocol used was; one integrated hermaphrodite was initially crossed with male wild-type worms and male transgenic progeny picked for cross two with wild-type hermaphrodites. The third cross is between a transgenic hermaphrodites and wild-type males and male transgenic progeny picked for cross four. The final cross is between transgenic males and a single wild-type hermaphrodite, this should give a fully integrated line (Figure 2.1). Approximately 20 transgenic hermaphrodites were picked from the last cross onto individual seeded NGM plates, and 100% of the progeny were transgenic worms.



### 2.1.5 *C. elegans* RNA interference (RNAi)

The RNAi method used in *C. elegans* has been adapted from WormBook ([http://www.wormbook.org/chapters/www\\_introreversegenetics/introreversegenetics.html](http://www.wormbook.org/chapters/www_introreversegenetics/introreversegenetics.html)). For this study RNAi was carried out using the feeding method. IPTG agar plates (Appendix 1) were seeded with a bacterial strain containing a plasmid expressing dsRNA from two T7 promoters, and incubated at room temperature overnight to enable dsRNA production. Two different methods were used; the first involved ~20 adult females being picked onto an un-seeded NGM agar plate and left for approximately 5 min, allowing time for the gut to clear of bacteria. These adults were then transferred onto one RNAi plate and allowed to egg lay for approximately 2 hours. 50-70 eggs per plate was a sufficient number and the adults were subsequently removed. It takes approximately three days for the worms to reach a suitable age for analysis. The second method involved transferring L3/L4 stage larvae onto the RNAi plates, after a 5 min period on an un-seeded OP50 plate to clear the gut. This method is favoured if the RNAi effect is likely to affect development, as the RNAi effects on development are likely to be less detrimental in the already developed larvae.

The RNAi feeding method was also carried out in liquid culture and on NGM agar plates at a laboratory at Pfizer. In the previously mentioned RNAi feeding method, worms were maintained on IPTG agar plates, with the IPTG enabling the production of dsRNA, thus an alternative method for dsRNA production in culture was required. One colony of the plasmid in *E. coli* bacteria was cultured overnight in a shaker at 37°C in 50 ml of L. Broth (Appendix 1). 0.4 mM IPTG was added to L. Broth (Ongvarrasoponea *et al.*, 2007) and was incubated for an additional 4 hours at 37°C before being used to feed mixed stage worms. Alternatively, the bacteria, L. broth and IPTG solution were used to seed NGM agar plates which were incubated at 37°C for 4 hours before mixed stage worms were added. The latter method proved more effective.

## 2.1.6 Antibody localisation methods

### 2.1.6.1 Generation of an anti-peptide antibody

Anti-peptide antibodies were generated at CovalAb in France against the *C. elegans* CPR-6 protein. Two *C. elegans* peptide regions that were highly conserved in *H. contortus* CPR-6 were identified. A BLAST search was carried out against the *C. elegans* and *H. contortus* genome sequences to ensure that these regions were unique. Approximately 26 mg of the Peptide regions **SFDSRDNWPKCDSIKV** and **PHDLYPTPKCEKKCV** were synthesised and conjugated at Covalab. These were then used to immunise two rabbits with both peptides. 0.5 ml of antigen and 0.5 ml of Freund's adjuvant were injected on days 0, 21, 42 and 63. An initial pre-immune bleed was taken on day 0 and post-immunisation bleeds on days 53 and 74. At each test bleed an Enzyme-linked immunosorbent assay (ELISA) was carried out to monitor the immunoreactivity of the antisera. A final bleed was taken on day 88 from which anti-peptide IgG was subsequently immunopurified using an affinity column. The final antibody concentration obtained was 9.2 µg/ml.

### 2.1.6.2 Preparation of protein extracts

Approximately 1,000 *H. contortus* ex-sheathed (Section 2.2.1) L3 larvae were centrifuged and washed in M9 buffer with the supernatant removed to leave a final volume of approximately 30 µl. *C. elegans* wild-type worms from a plate

that had just cleared of bacteria were washed off in M9 buffer, centrifuged and the supernatant removed to leave a final volume of approximately 30  $\mu$ l. 30  $\mu$ l of SDS loading buffer were added to this and heated at 95°C for 10 min. This buffer is 2x SDS-PAGE buffer (Appendix 1) and 0.7 M 2-mercaptoethanol (Sigma-Aldrich). The mixture was stable enough to be stored at -20°C until required.

### **2.1.6.3 Protein separation by polyacrylamide gel electrophoresis**

Protein extracts were separated using sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE). Approximately 20  $\mu$ l of previously prepared sample (section 2.1.6.2) was heated to 95°C for a few min before use. Samples were loaded onto a 4-15% Tris-HCl Gel (BioRad) and run in 1x Tris/Glycine/SDS running buffer (BioRad) (Appendix 1) at 200 V for 1 hour using a BioRad Power Pac 200. A pre-stained protein marker was included to allow sample visualisation.

### **2.1.6.4 Western Blotting of Ce-CPR-6**

The separated protein samples were then transferred onto an activated Hybond-P membrane (Amersham). A sandwich consisting of (from back to front); a pre-soaked sponge, filter paper, gel, membrane, filter paper and sponge was set-up for the transfer. The transfer was carried out in a chamber filled with chilled 1x Tris/Glycine blotting buffer (BioRad) (Appendix 1) and an ice box and was run at 100 V for 1 hour. The membrane was removed from the sandwich and rinsed in 1x PBST (Appendix 1) and then left to shake for 20 min in 5% block solution (Appendix 1). The block was removed and a 1/2000 dilution of the primary CPR-6 antibody (CovalAb) in 5% block solution added. This was left to shake at 4°C overnight. The sample was washed three times with 5% block solution to remove the primary antibody and a 1/10000 dilution of purified goat anti-rabbit IgG HRP antibody in 5% block solution added. This was left to shake at room temperature for 1 hour. The sample was then washed three times in 1x PBST. The ECL Plus Western Blotting Detection System (Amersham) was used for antibody detection following the manufacturer's instruction. The kit produces a fluorescent light emission that can be detected by exposure to x-ray.



### 2.1.6.5 Freeze-crack method for worm immunofluorescence

A glass slide was coated with approximately 100  $\mu\text{l}$  of poly-l-lysine (Sigma-Aldrich) in a square in the centre. Approximately 10  $\mu\text{l}$  of water was added to the poly-l-lysine coated slide and a number of worms (20+) picked into this and using a scalpel blade the individual worms were cut behind the head to expose the gut. A glass cover slip was placed over the worms and the slide put on a metal plate within a dry ice filled box, and placed at  $-80^{\circ}\text{C}$  for at least 30 min. The cover slip was cracked off using the end of a scalpel blade and the worms fixed by immersion in methanol at  $-20^{\circ}\text{C}$  for 10 min and then in acetone at  $-20^{\circ}\text{C}$  for 10 min. Slides were stored at  $-20^{\circ}\text{C}$  until required.

### 2.1.6.6 CPR-6 antibody staining method

The slides prepared in section 2.1.6.5 were incubated in 1x PBST for a few min then 1% block solution (Appendix 1) added. After 20 min, the block was gently tipped off and a 1/50 dilution of CPR-6 antibody in 1% block solution added to the slide. The slide was placed in a humidifying chamber at  $4^{\circ}\text{C}$  overnight. The slide was gently washed three times in 1% block and a 1/200 dilution of a purified goat anti-rabbit FITC antibody in 1% block solution added, and placed in a humidifying chamber for 1 hour at room temperature. The slide was washed twice in 1x PBST and then rinsed with PBS. A 1/1000 dilution of DAPI (Invitrogen) was added to the slide and left for 1 min before rinsing with PBS; this was to aid identification of the worms under the microscope. A 1/2 dilution of Prolong Gold antifade reagent (Invitrogen) in water was added to the slide to prevent fluorescence fading during imaging before a cover slip was applied. Worms were examined using an Axioscop 2 plus microscope (Zeiss), excitation/emission for DAPI  $\sim 350/470$  nm and for FITC  $\sim 485/535$  nm.

### 2.1.7 *C. elegans* drug screening

Screening of transgenic *C. elegans* CLB01 worms (chapters 2.1.4 and 5.2.1.4) was carried out at Pfizer Animal Health. As Pfizer do not routinely use *C. elegans* worms, it was necessary to establish all conditions for drug assays. The specific details of how the screen was generated and carried out are found in Chapter 6 results section 6.2.

## **2.2 *H. contortus* methods**

### **2.2.1 *H. contortus* larval culture**

All the *H. contortus* used in this study were ISE strain unless otherwise stated. *H. contortus* larvae were kept in water in flasks at 4°C until they were required for use. Approximately 1,000 larvae were transferred into a 1.5 ml eppendorf, centrifuged pelleted and the majority of the supernatant removed. Larvae were exsheathed by adding 200 µl of 1x Phosphate Buffered Saline (PBS) (Sigma-Aldrich) (Appendix 1) and 5 µl of sodium hypochlorite solution (10-13%, Sigma-Aldrich) and left for 5 min. The larvae were then washed three times with 1x PBS to remove all the sodium hypochlorite. In a sterile hood the larvae were washed twice in Earle's Balanced Salt Solution (EBSS) pH 5.2 (Invitrogen) containing Fungizone (1.25 µg/ml), Penicillin (250 units/ml) and Streptomycin (50 µg/ml). After the final wash and spin, the pellet was re-suspended in 200 µl of EBSS (containing Penicillin/Streptomycin and Fungizone) and incubated at 37°C for 3-7 days. Cultured *H. contortus* larvae were used for cDNA synthesis and examination of CPR-6 antibody localisation patterns.

### **2.2.2 *H. contortus* drug screening**

A standard plate of 96 or 384 wells requires 8,000 viable larvae per ml. 10 ml of *H. contortus* L3 larvae (USA strain), from water flasks at 4°C, were centrifuged for 3 min at 1,000 g and the supernatant removed. The larvae were then washed in water and the supernatant removed. The volume was made up to 90 ml with Glucose Tyrodes balanced salt solution (Appendix 1). Before screening, larvae were ex-sheathed using 10 ml of 2% Sodium Hypochlorite solution. This mix was centrifuged for 3 min at 1,000 g and the supernatant removed. 50 ml of Glucose Tyrodes balanced salt solution was added and the mixture centrifuged for a further 3 min at 1,000 g, following which the supernatant was removed leaving an ex-sheathed larval pellet. Glucose Tyrodes balanced salt solution washes were repeated three times to ensure all the bleach was removed. The larvae were re-suspended in Basal media (Appendix 1) and Penicillin, Streptomycin and Gentamicin added. A volume of 100 µl and 25 µl per well were added to 96 and

380 well plates, respectively, giving approximately 800 larvae per well in 96 well plates and about 200 larvae per well in 380 well plates.

## **2.3 Molecular biology methods**

### **2.3.1 Polymerase Chain Reaction (PCR)**

#### **2.3.1.1 Standard PCR protocol**

Polymerase Chain Reaction (PCR) was carried out using a Techne Flexigene PCR machine. GoTaq Flexi DNA polymerase (Promega) was used for standard reactions, prepared following the manufacturer's instructions (1x Green GoTaq Flexi Buffer, 1 mM MgCl<sub>2</sub>, 0.125 mM of each dNTP, 0.5 μM of forward and reverse primers, up to 1.25 units of GoTaq Flexi DNA polymerase and 50-100 ng of DNA template). Standard PCR conditions were: initial denaturing for 3 min at 95°C, 25-35 cycles of 30 sec at 95°C to denature, 30 sec at 55-57°C for primer annealing and 1 min per kb at 72°C for extension, followed by a final extension of 3 min at 72°C.

#### **2.3.1.2 Proof-reading PCR protocol for gene promoter amplification**

Where PCR products were required for promoter expression studies, the proofreading Taq *PfuUltra* II Fusion HS DNA polymerase (Stratagene) was used. Standard protocol was followed using the 10x *PfuUltra* II Reaction Buffer provided. PCR products were blunt ended and thus could not be cloned directly, therefore 3' adenosine overhangs were added by incubation of the PCR reaction mix (20 μl) for 5 min at 72°C with 1.25 units of GoTaq Flexi DNA polymerase and 0.1 mM dATP.

#### **2.3.1.3 Rapid Amplification of cDNA Ends (RACE)**

Rapid Amplification of cDNA Ends (RACE) was used to obtain the start of the *H. contortus* cysteine proteases present on BAC 18f22 and both the start and end of the *Hc-cpr-6* gene. RNA was extracted from adult stage *H. contortus* using the Qiagen RNeasy Mini Kit (QIAGEN). *H. contortus* cDNA was first prepared from 1.6 μg using the AffinityScript Multi Temperature cDNA Synthesis Kit (Stratagene)

and was used in the first round PCR. The FirstChoice RLM-RACE Kit (Ambion) was used to prepare the 5' and 3' RACE products and a nested PCR protocol was followed. The first round PCR protocol: 2 µl cDNA, 1 mM MgCl<sub>2</sub>, 0.125 mM of each dNTP, 0.5 µM of kit outer primer, 0.5 µM of gene specific outer primer, 1x Green GoTaq Flexi buffer and up to 1.25 units of GoTaq Flexi DNA polymerase. PCR conditions were: denaturing for 3 min at 95°C, 35 cycles of 15 sec at 95°C, 15 sec at 57°C and 1 min at 72°C, followed by 15 sec at 95°C, 15 sec at 57°C and 3 min at 72°C.

The second round PCR used 1 µl of the first round reaction as a template. The PCR cycle was the same as the first round and standard PCR reaction as above but replacing first round primers with 0.5 µM of the kit inner primer and 0.5 µM of the gene specific inner primer.

3' RACE was used to obtain the end of the *H. contortus cpr-6* gene using the same kit. *H. contortus* RNA, 0.03 µg 3' RACE Adapter primer and 14.2 µl of RNase free water were incubated at 65°C for 5 min then cooled to room temperature for 10 min. Added to this were 1x RT Buffer, 0.02 µM dNTP mix, 20 units of RNase Inhibitor and 20 units of Reverse Transcriptase. This mixture was incubated at 42°C for 60 min and then at 72°C for 10 min, at this point the cDNA was stable and could be stored at -20°C. Standard PCR was carried out using 2 µl of the generated cDNA, 1x Green GoTaq Buffer, 1 mM MgCl<sub>2</sub>, 0.125 mM of each dNTP, 0.5 µM of 3' RACE kit outer primer, 0.5 µM of gene specific primer and up to 1.25 units of GoTaq Flexi DNA polymerase in a total volume of 30 µl. PCR cycles were as for 5' RACE.

#### **2.3.1.4 Worm lysis PCR**

Worm lysis PCR was used for confirming the presence of specific genes and marker genes in the *C. elegans* worm. 6-10 *C. elegans* adult worms were added to a standard PCR tube containing 4 µl of a Single Worm Lysis Buffer (SWLB) (Appendix 1) and 0.1 mg proteinase K. Tubes were placed at -80°C for at least 30 min, incubated at 60°C for 1 hour to lyse the worms and then placed at 95°C for 15 min. The standard PCR protocol was followed using the 4 µl solution from above as the DNA template and gene specific primers.

### 2.3.1.5 Reverse Transcription PCR

Reverse Transcription PCR was used to check the presence of expressed genes in *C. elegans* and *H. contortus* (section 2.3.1.3) cDNA. Approximately 80 adult *C. elegans* worms were picked into M9 buffer and re-suspended in 200 µl of lysis buffer containing 0.1 mg of Proteinase K and 0.7 M of 2-mercaptoethanol. The sample was incubated at -80°C for at least 30 min and heated to 55°C for one hour, allowed to cool and incubated at 4°C for 10 min. 500 µl of Total RNA Isolation Reagent (ABgene) was added and mixed by vortexing then incubated at 4°C for 10 min. 100 µl of CHCl<sub>3</sub> was added, mixed well, then incubated on ice for 5 min and then centrifuged at 13,000 g for 15 min. The top phase was collected and precipitated in an equal volume of Isopropanol. This was then incubated on ice for 10 min and centrifuged at 13,000 g for 10 min, followed by a wash with 70% ethanol and centrifuged for 5 min at 13,000 g. The supernatant was removed and the pellet air dried and re-suspended in 30 µl of RNase-Free Water (QIAGEN). The SuperScript One-step RT-PCR System (Invitrogen) was used to produce cDNA from the *C. elegans* RNA and for the PCR reaction. The PCR conditions were: 30 min at 50°C and 2 min at 94°C, then 30 cycles of 30 sec at 94°C, 30 sec at 58°C and 30 sec at 72°C, then a final extension of 3 min at 72°C.

### 2.3.2 Agarose gel electrophoresis

Standard gels used for viewing PCR products were 1-2% UltraPure Agarose (Invitrogen) dissolved in 1x Tris-acetate EDTA (TAE) electrophoresis buffer (Appendix 1). Products were visualised using 0.5 µg/ml Ethidium bromide or 0.05 µl/ml SafeView Nucleic Acid Stain (NBS Biologicals). PCR products were mixed with DNA loading buffer (Appendix 1) and loaded onto gels. Electrophoresis was performed using Gibco BRL Horizontal Gel Electrophoresis Apparatus at 100-120 V in 1x TAE. Visualisation was with BioRad Trans UV Illuminator (BioRad). Product sizes were identified using 100 bp and 1 kb DNA ladders (Invitrogen) (Appendix 1).

### 2.3.3 Purification of PCR products

The PCR product was viewed on the gel using a High Performance UV Transilluminator (UVP) which has a lower UV setting and is thus less likely to

cause mutations to the DNA. The DNA band was excised using a scalpel and placed in a 1.5 ml eppendorf. Gel extraction was then carried out using the QIAquick Gel Extraction Kit (QIAGEN).

### 2.3.4 Restriction enzyme analysis

Restriction enzyme digestion was carried out both to confirm the presence of a specific sequence and to extract an insert from a vector, e.g. for cloning. Restriction endonucleases (New England Biolabs) were used and the standard protocol with specific buffers was followed.

Vectors that were to be used for cloning were treated with Alkaline Phosphatase (New England Biolabs) post digestion. 1 unit of Alkaline Phosphatase was added to the digest and incubated at 37°C for 30 minutes; this treatment removes the 5' phosphate group and prevents the vector from self-ligating.

Products were viewed using the method described for PCR products in Section 2.3.2 and purified using the technique described in Section 2.3.3.

### 2.3.5 Cloning into pCR 2.1-TOPO

The TOPO 2.1 TA Cloning Kit (Invitrogen) was used for one step cloning of PCR products amplified by Taq. Standard protocol was followed and 2 µl of the cloning reaction added to TOP 10 one shot ultracompetent cells (Invitrogen) for transformation. 400 µl of SOC medium (Appendix 1) was used, with 120 µl of the transformation reaction being spread on an L. Agar plate containing Ampicillin (Appendix 1). Blue/white selection of colonies containing PCR products was utilised by the addition of 50 µl of IPTG and 50 µl of 2% X-Gal (Appendix 1) to the Ampicillin plate and placed at 37°C overnight.

### 2.3.6 Cloning into Fire lab vectors

The Fire lab vectors (Fire *et al.*, 1990) have been used for studying gene expression and function of both *C. elegans* genes and *H. contortus* expressed genes in *C. elegans*. Initially a ligation between the promoter and vector sequence using T4 DNA Ligase (New England Biolabs) was carried out. A 3:1 ratio

of insert to vector was desired, aiming for 100 ng of insert per ligation reaction. The standard ligation reaction consisted of 1 x ligation buffer, 1 unit ligase, 100 ng insert and vector in water. This was incubated at 4°C overnight.

Transformation was carried out using XL10-Gold ultracompetent cells (Stratagene). 2 µl of the ligation reaction was added to 40 µl of thawed cells and placed on ice for 30-60 min. The reaction was then heat shocked at 42°C for 30 sec, 400 µl of SOC medium was added and then left to shake for 1 hour at 37°C. 120 µl of this was then spread onto an Ampicillin L. Agar plate and incubated at 37°C overnight.

### **2.3.7 Identification of positive cultures**

For cloning into the Fire lab vectors no selection is used and any colony may be positive. A single colony was lysed in 100 µl of water and heated to 95°C for 10 min. 5 µl of this lysed colony preparation was used as the DNA template for PCR screening. Standard PCR was carried out using either vector or gene specific primers. Any positive colonies were cultured in 10 ml of L. Broth with 0.1 µg of Ampicillin overnight in a shaker at 37°C.

### **2.3.8 DNA purification**

For the purification of TOPO cloned colonies grown in L. Broth and Ampicillin, purification of up to 20 µg of DNA was carried out using the Plasmid Mini Kit (QIAGEN) as per manufacturer's instruction. Purification of Fire lab vectors for subsequent microinjection into *C. elegans* worms was carried out using the QIAprep Spin Miniprep kit (QIAGEN) as per manufacturer's instruction.

### **2.3.9 Preparation of compounds for drug screening**

Compounds supplied as powders were solubilised in Dimethyl Sulfoxide (DMSO) using a Tecan Freedom Evo giving a final volume of about 100-200 µl, depending on the molecular weight of the compound. A stock concentration of 30 mM in 100% DMSO for each compound was achieved. A final working percentage of below 0.5% DMSO is required when working with *C. elegans*, to ensure that there is no DMSO interference during drug screening.

## 2.4 Bioinformatic methods

### 2.4.1 Software and databases used

Vector NTI Advance Software (Invitrogen, versions 10 & 11) was used for sequence analysis of *C. elegans* and *H. contortus* sequences. *C. elegans* sequences are available from wormbase (<http://www.wormbase.org/>) and *H. contortus* sequences from NCBI (<http://www.ncbi.nlm.nih.gov/>) and the *H. contortus* BLAST server on the Sanger web page ([http://www.sanger.ac.uk/cgi-bin/blast/submitblast/h\\_contortus](http://www.sanger.ac.uk/cgi-bin/blast/submitblast/h_contortus)). Alignment and direct sequence comparison was carried out using Align X which is part of the Vector NTI software and the ClustalW2- Multiple Sequence Alignment tool, available on the European Bioinformatics Institute website (<http://www.ebi.ac.uk/Tools/msa/clustalw2/>). Gene sequence analysis was carried out using the Artemis Genome Browser and Annotation Tool which is available to download from the Sanger website (<http://www.sanger.ac.uk/resources/software/artemis/>). The MEGA (Molecular Evolutionary Genetic Analysis) Tool version 5 (<http://www.megasoftware.net/>) was used for phylogenetic analysis. A number of the figures contained in this thesis were generated using Microsoft Office PowerPoint and Microsoft Office Excel. Sequence LOGOS within gene promoters were identified using MEME suite, which is a motif based sequence analysis tool (<http://meme.sdsc.edu/meme/cgi-bin/meme.cgi>).

### 2.4.2 DNA primers for PCR

Primers were ordered from the Eurofins MWG Operon website (<http://www.eurofinsdna.com/home.html?gclid=COafvvyhpcCFQ9hQgodLQnn9g>). 1x TE (Appendix 1) was added to give a stock concentration of 1 µg/µl and primers were stored at -20°C. A working stock solution of 50 ng/µl was obtained by diluting with water.

### 2.4.3 DNA sequencing

Samples were sent to Eurofins MWG Operon for sequencing. This was carried out after insertion of the desired DNA sequence into a vector, most commonly TOPO.



Purified plasmid DNA was supplied at a concentration of 50-100 ng/ $\mu$ l in a minimum whole volume of 15  $\mu$ l of double distilled water. Either vector primers or gene specific primers were used to carry out sequencing.

Sequencing results were analysed in Contig Express which is part of the Vector NTI software. The results are displayed in both word and chromatogram form. The chromatogram is useful as it depicts the ratio of each of the four bases found at a specific location during sequencing. Analysis can continue in Align X or ClustalW2.

#### **2.4.4 Signal peptide cleavage identification**

The Signal P programme (<http://www.cbs.dtu.dk/services/SignalP/>) was used to identify the signal peptide region of proteins. This region is usually between 3 and 60 amino acids in length and is involved in protein transportation from the endoplasmic reticulum to the Golgi apparatus for further processing of proteins. There are three different scores provided in the output information of this programme C, S and Y. The S score is the predicted signal peptide, taking into account every amino acid position. High S scores indicate that a particular amino acid is part of a signal peptide and a low S score indicative of being part of a mature protein. The C score is the cleavage site and is highlighted between two amino acids. The second amino acid is the start of the mature protein. The Y score is a cleavage site prediction based on a combination of the C score and S score. This may give a more accurate identification of the true cleavage site as usually there are multiple high scores within a sequence.

#### **2.4.5 Computational gel analysis for expression levels**

Image J is available to download and was used for the comparison of PCR products on agarose gels, e.g. *cpr-6* expression in *H. contortus* (<http://rsbweb.nih.gov/ij/download.html>). Primers for both the gene of interest and a control gene e.g. *sod-1* were added to the same PCR reaction. On observation of the subsequent gel there were two bands in each lane. The gel image was used for analysis in Image J, comparing the intensity of the two bands.

## **Chapter 3**

Annotation and characterisation of a cathepsin B  
cysteine protease gene family present on *H.*  
*contortus* BAC 18f22

### 3.1 Introduction

The genome of *H. contortus* is currently being sequenced, with completion anticipated 2012/2013. Current data suggests that the *H. contortus* genome may be over four times the size of original estimates (300-400 Mb). This size is estimated using the N50 statistic which can be applied to contiguous sequence databases. N50 is recorded in kb and is widely used in genome size estimation and assembly. Initially, contigs are arranged in size from largest to smallest, and the lengths added together. The N50 is the size of the last sequence that results in the summed length exceeding half of the total length of the database (Miller *et al.*, 2010). The predicted size may be larger than original estimates due to the increased number and size of introns in the *H. contortus* genome compared to *C. elegans* (Laing *et al.*, 2011). This genome data is available as contigs, supercontigs and as Bacterial Artificial Chromosomes (BACs). Supercontigs are individually sequenced regions that have been assembled from overlapping sequence reads, whereas BAC reads are more reliable as they have been sequenced in full from a single clone. Analysis carried out during this project made use of the Sanger databases; *H. contortus* assembled supercontigs (21/08/08)(all reads), *H. contortus* supercontigs (26/08/2009), *H. contortus* assembled BAC contigs and the most recent assembly available on the Sanger ftp site (12062012). From this latest assembly (12062012) the N50 is 83,238 bp (James Cotton, personal communication).

Although a significant amount of sequence information has been generated for *H. contortus*, reliable assembly, which is required for gene annotation, has proved a major hurdle. This is thought to be due to heterogeneity of the parasite population, with variation in gene sequence making it difficult to identify overlapping reads at high stringency. It is hoped that increased depth of sequencing combined with sequence from more inbred strains will overcome this.

Due to the blood-feeding activity of *H. contortus* adult worms, there has been great interest in enzymes present within the gut and their putative role in blood digestion and anticoagulation. Previous studies have identified a number of protease gene sequences from *H. contortus* that encode cathepsin B cysteine

proteases (Rehman and Jasmer, 1999; Skuce *et al.*, 1999). Work by Cox *et al.* (1990) first identified that protein extracts from adult worms can degrade fibrinogen and subsequently increase clotting time. One of the major components of these extracts was a 35 kDa cysteine protease. This protein was termed AC-1 and was identified as being capable of degrading fibrinogen *in vitro* and may be a potential anticoagulant (Cox *et al.*, 1990). AC-1 was shown to have significant homology (42% identity) with mammalian cathepsin B when compared to sequences available for human, rat and mouse. Other proteases with high sequence similarity to AC-1 were subsequently identified, resulting in the characterisation of a Cathepsin B-like (CBL) family containing five genes (AC-1, AC-2, AC-3, AC-4 and AC-5) (Pratt *et al.*, 1992).

AC-like proteases were the first to be identified in *H. contortus* and subsequently other cathepsin B proteases have been identified from US strains (Cox *et al.*, 1990; Pratt *et al.*, 1990; Pratt *et al.*, 1992). To date, their presence has to be confirmed in other strains. In addition it was unclear whether these represented distinct genes or population polymorphisms. Most of these are distinct from proteases in *C. elegans* (Pratt *et al.*, 1992), suggesting roles specific to the parasite, most likely in anticoagulation and blood digestion (Baig *et al.*, 2006). Work has been carried out looking at CBL cysteine proteases in other blood feeding parasites, namely hookworms *Ancylostoma caninum* and *Necator americanus* (Ranjit *et al.*, 2008).

At the start of this project a *H. contortus* BAC was identified which contains sequence encoding several cathepsin B proteases.

The main aims of this chapter were to;

- Identify and annotate the *H. contortus* cathepsin B protease genes present on the BAC.
- Examine and compare the sequence, structure and expression of the cathepsin B cysteine protease genes on the *H. contortus* BAC.
- Determine the relationship of the encoded proteases to the previously identified cathepsin B enzymes for *H. contortus* and other nematodes.

## 3.2 Results

### 3.2.1 Protease genes on *H. contortus* BAC 18f22

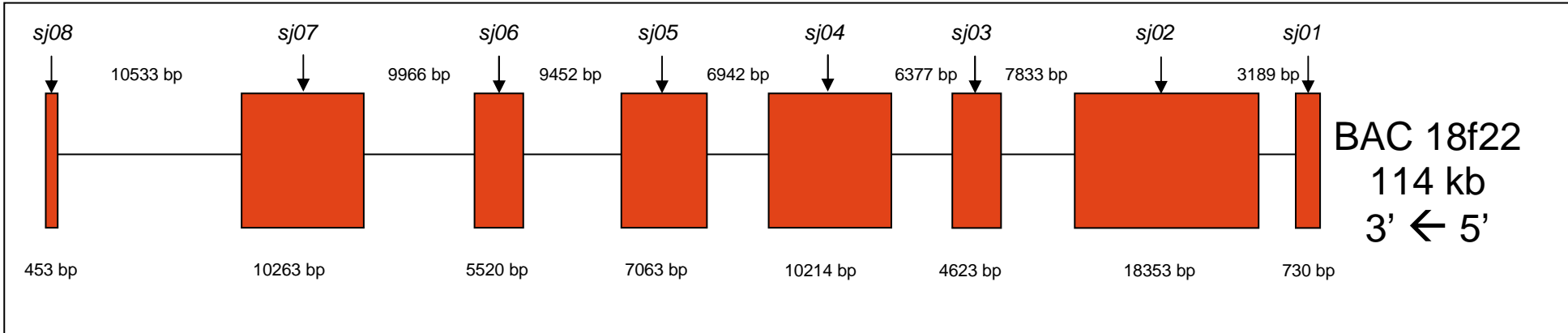
#### 3.2.1.1 Annotation of genes on *H. contortus* BAC 18f22

Bioinformatic analysis identified eight genes encoding cysteine proteases present on *H. contortus* BAC 18f22. This BAC is 113,944 base pairs in length and the eight genes are tandemly arranged, spanning the entire length of the BAC. These genes were identified by tBLASTn analysis of *H. contortus* assembled BAC contig sequences ([http://www.sanger.ac.uk/cgi-bin/blast/submitblast/h\\_contortus](http://www.sanger.ac.uk/cgi-bin/blast/submitblast/h_contortus)) using the previously identified *H. contortus* AC-2 protein. The genes are therefore related to the *H. contortus* AC cysteine protease gene family. The genes identified were numbered *sj01-08*, from right to left (due to their location on the negative strand) along the BAC. The position on the BAC and size of each gene is shown in Figure 3.1. *sj01* is incomplete due to its location near the 5' end of the BAC, with only the last three exons present. *sj08* is located near the 3' end of the BAC, resulting in only the first three exons being present. Further work searching for sequences overlapping with BAC 18f22 identified the rest of the *sj08* gene sequence at a different genomic location, on a supercontig (Section 3.2.1.5). The start methionine could not be identified for any of the genes so 5' RACE was performed to attempt to confirm the location of the first exon of each gene (Section 3.2.1.2).

Using both BLAST (with the *H. contortus* AC-2 sequence) and manual analysis, most of the genes could be almost fully annotated using the Artemis Genome Browser and Annotation Tool which is available to download from the Sanger website (<http://www.sanger.ac.uk/resources/software/artemis/>). As mentioned previously, some parasite genomes are already fully annotated. This however is not the case for *H. contortus*. Annotation of the *H. contortus* BAC was carried out to obtain the location and structure of the genes present on the BAC. Figure 3.2 is an Artemis screenshot, indicating the location and structure of each gene on the BAC. In Artemis, exons are indicated by solid boxes and introns indicated by lines joining these boxes. Initial analysis of the structure of each gene showed

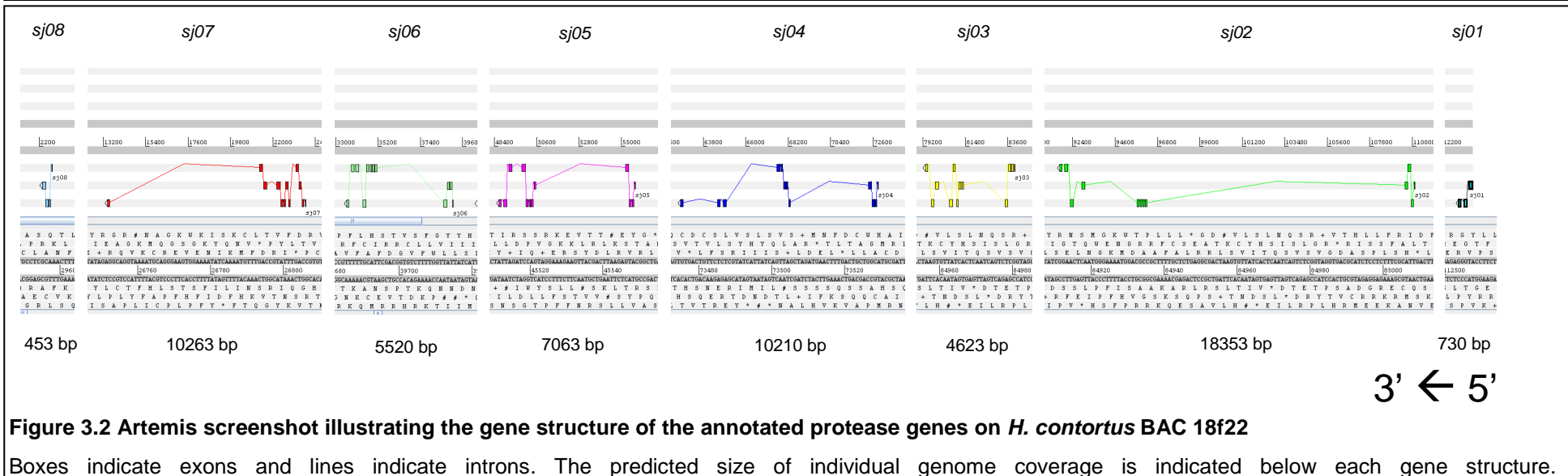
that while exons are identical in length, intron sizes are more variable. This is reflected in the large variations in gene size (4,623-18,353 bp).

All of the genes which could be fully annotated (*sj02-sj05*) contained twelve exons, however the last exon for *sj06* and *sj07* could not be identified using the AC-2 sequence. This suggested that the last exon of both genes was present, but was too divergent to be identified by AC-2 BLAST analysis. These exons were subsequently identified by carrying out a tBLASTn search using the C' terminal region of the rat cathepsin B sequence (rcatB, P00787) (Guenette *et al.*, 1994).



**Figure 3.1 Diagrammatic representation of the size and approximate location of genes on *H. contortus* BAC 18f22**

The middle line along the length of the diagram indicates the 114 kb BAC 18f22. Genes are present on the negative strand running from right to left. Red boxes represent gene locations with the predicted size of their genome coverage indicated below, and the intergenic regions indicated above the boxes. *sj01* and *sj08* are incomplete due to their location near the start and end of the BAC, respectively.



### 3.2.1.2 Identification of the 5' end of the cysteine protease genes on *H. contortus* BAC 18f22

tBLASTn analysis failed to identify the start codon for the genes present on the BAC. Previous work found that for AC-2 and AC-3 the first exon comprised only the start methionine (Pratt *et al.*, 1990). To identify the first exon of the BAC genes, 5' RACE was carried out. This identified the location of the start codon for *sj03*, *sj04*, *sj05*, *sj06* and *sj07* (primer sequences in Appendix 2 Table 3.1). Alignment of the 5' RACE products with the genomic sequence for each gene showed that the first exon encoded only the start methionine.

Although 5' RACE products for *sj03-sj07* could be aligned to the BAC sequence, the sequences were not identical. The 5' RACE products spanned the first three exons (150 bp) and within this region there were a number of base pair changes compared to the BAC sequence (11 in *sj03*, 4 in *sj04*, 6 in *sj05*, 13 in *sj06* and 9 in *sj07*). For all of the genes except *sj06* the majority of these changes occurred in exon 2. *sj06* however had 13 bp changes within the third exon. The 5' UTR sequence was well conserved. There were two base pair changes in *sj05* and one in *sj07*, but as these regions are non-coding this is not significant. On closer examination at the amino acid level, there does not appear to be a pattern as to whether base pair changes are silent or result in an amino acid change, with a proportion of both being varied across the genes. These differences could be due to variation between worms and highlights the heterogeneity in the parasite population (Prichard, 2001). Alternatively these differences could be due to amplification of closely related but not identical protease genes. The BAC represents only one genomic sequence for each AC-related gene and therefore variation from this was anticipated. For all sequence analysis carried out, the original BAC sequence was used. Few genes of *H. contortus* have been annotated at the genomic level and compared to cDNAs, but similar analysis of the *H. contortus* cathepsin L gene *Hc-cpl-1* also found up to eight single nucleotide polymorphisms (SNPs), most of which were silent substitutions, over the 1,062 bp gene (Britton and Murray, 2002).

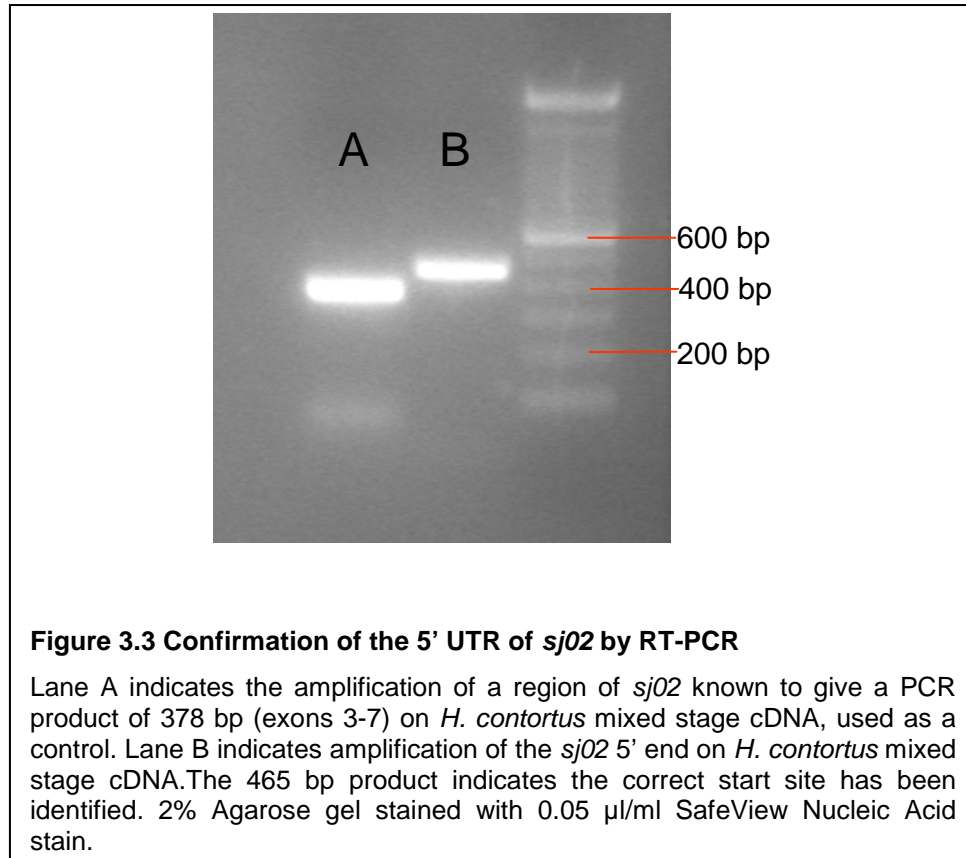
Only the last three exons are available on the BAC for *sj01* and thus 5' RACE was not attempted for this gene. The start codons for the *sj02* and *sj08* genes could not be confirmed by 5' RACE despite the more thorough nested 5' RACE PCR



protocol followed. It was thus speculated that the lack of RACE product could be due to low expression of these genes, or that the genes were pseudogenes.

To confirm that *sj02* is expressed, RT-PCR was carried out on *H. contortus* cDNA using primers designed across exons 3 to 7 (primer sequences in Appendix 2 Table 3.2). A product of the desired size (378 bp) was obtained, confirming that *sj02* is expressed. As 5' RACE proved unsuccessful after numerous attempts with varied conditions, an alternative approach to obtain the start codon was attempted. In *C. elegans* a number of genes have been shown to be trans-spliced, resulting in the 5' end of a transcript being replaced by a splice leader (SL) sequence, reviewed in Allen *et al.*, (2011). The exact role that trans-splicing plays in *C. elegans* is still unknown, however it may have a role in translation initiation (Lall *et al.*, 2004). There are two SL sequences in *H. contortus*, SL1 (GGTTTAATTACCCAAGTTTGAG) and SL2 (GGTTTAACCCAGTTACTCAAG) (Bektesh *et al.*, 1988; Blumenthal, 2005). PCRs using either SL1 or SL2 as the forward primer were carried out on cDNA with the *sj02* reverse primer designed to exon 7 (primer sequence in Appendix 2 Table 3.2). Neither produced a product, indicating that *sj02* is not trans-spliced.

It was however possible to predict the location of the first exon for *sj02* and *sj08* by examining the genomic information available for the BAC. For genes *sj03-sj07* the confirmed size of the first intron was in the range of 58 to 96 bp in length. Taking this range of intron size into consideration a proposed start codon for both the *sj02* and *sj08* genes was identified. For *sj02*, PCR was carried out using a primer incorporating this sequence (5' UTR and start codon) together with a reverse primer designed to the seventh exon (primer sequences in Appendix 2 Table 3.3). This produced a product of 465 bp on cDNA (Figure 3.3) and the sequence was confirmed by sequencing and mapped to the BAC sequence.



### 3.2.1.3 Intronic analysis of the *H. contortus* genes on BAC 18f22

Previous work in a number of species suggested that in addition to the upstream promoter region, the first intron may have an important role in regulating gene expression and for this reason may be a valuable region to examine (Bruhat *et al.*, 1990; Jeong *et al.*, 2006). Initial analysis of the non-coding regions of the BAC protease genes focussed on the first intron. While these were reasonably conserved in size (58-96 bp), no obvious long stretches of homology were found between the genes. One six base pair motif ATTGAA was identified in a number of the first introns and on subsequent analysis of other introns was found to be present in these also, which may or may not be significant. The highest identity was found between genes located consecutively on the BAC, for example *sj03* & *sj04* and *sj04* & *sj05* (Table 3.1), with high conservation of base pairs shown in Figure 3.4. This reflects their close positioning in the genome and suggests recent gene duplication. This is also suggested by conservation of the positions of all introns in all of the BAC protease genes (Figure 3.5). Boundary splice site sequences appear to be conserved between *H. contortus* and *C. elegans* (GU-AG)

(Blumenthal and Steward, 1997). Identification of the conserved sequence around the intron/exon boundaries aided gene annotation.

| Gene        | <i>sj02</i> | <i>sj03</i> | <i>sj04</i> | <i>sj05</i> | <i>sj06</i> | <i>sj07</i> | <i>sj08</i> |
|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| <i>sj02</i> | 100         | 57          | 58          | 55          | 55          | 52          | 56          |
| <i>sj03</i> |             | 100         | 72          | 61          | 54          | 48          | 40          |
| <i>sj04</i> |             |             | 100         | 70          | 57          | 59          | 56          |
| <i>sj05</i> |             |             |             | 100         | 55          | 52          | 61          |
| <i>sj06</i> |             |             |             |             | 100         | 57          | 59          |
| <i>sj07</i> |             |             |             |             |             | 100         | 65          |
| <i>sj08</i> |             |             |             |             |             |             | 100         |

**Table 3.1 DNA comparison of the first intron of the BAC protease genes**

The percentage identity of intronic DNA sequences was calculated from alignments. The highest identities, between *sj03* & *sj04* and between *sj04* & *sj05*, are indicated in red.

```

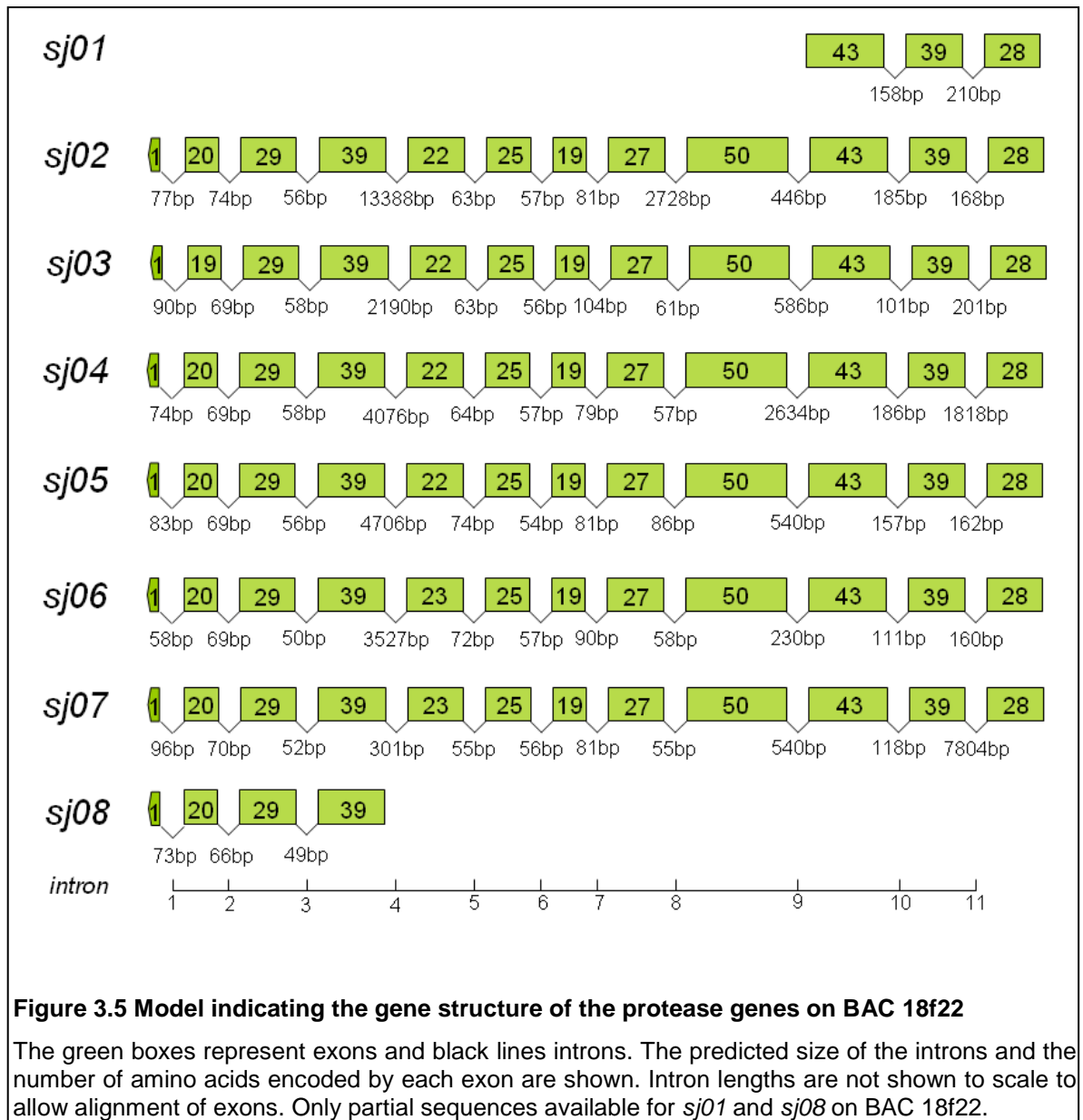
sj03      ATGGTTAGTTAACTGCCTGACCTACTTTTTAGAATAAATTTTTTCATTTATATTGAAGTAGTGAGGAATATTGAAAGCATATGTTTCAGATG 90
sj04      ATGGTAAGTTACCTCGCTGACCTTCTTTTACAATAATTTTTTCATTTGGATTGAA-----CGGCAGATTGAAAGATATATG----- 74

sj04      ATGGTAAGTTACCTCGCTGACCTTCTTTTACAATAATTTTTTCATTTGGATTGAACGGCAGATTGAAGATATATG----- 74
sj05      ATGGTGAGCTGGCTTGTGACCTTCTTTCTAAGAGGAGTTTTGACATTGAGATTGAACGACAGATCGAGAATAGTCGTTTCAGATG 83

```

**Figure 3.4 DNA alignment of the most conserved first introns of certain BAC genes**

The sequence runs from the start codon on the left (indicated in blue) with intron lengths indicated at the end. Base pair changes between the two pairs with highest identity are indicated in red.



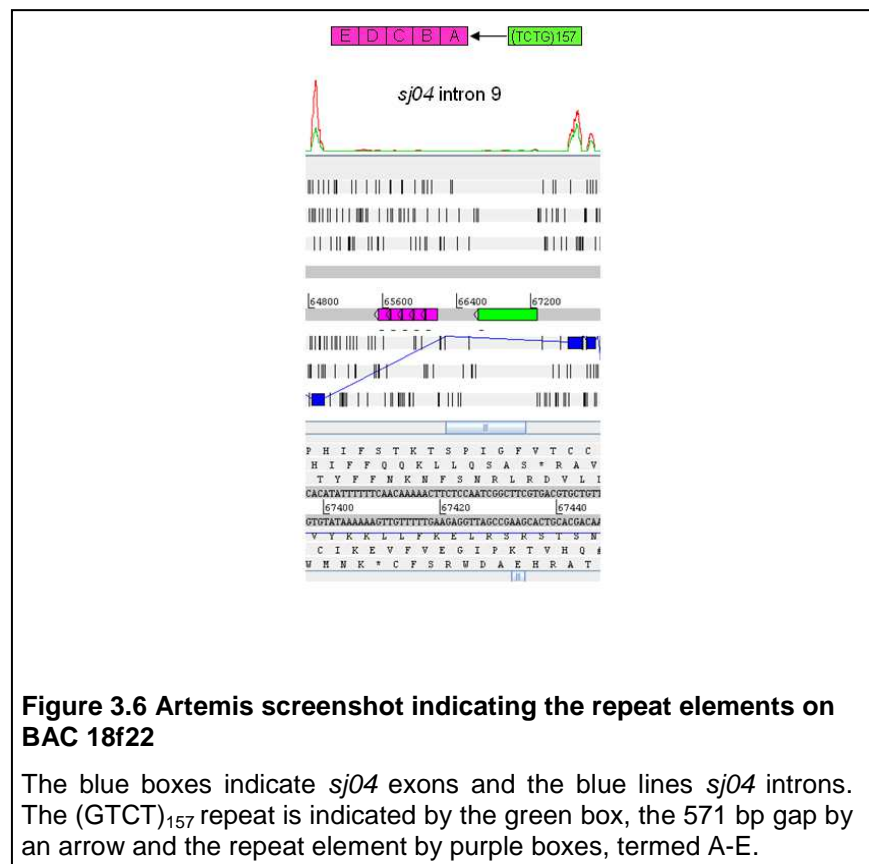
**Figure 3.5 Model indicating the gene structure of the protease genes on BAC 18f22**

The green boxes represent exons and black lines introns. The predicted size of the introns and the number of amino acids encoded by each exon are shown. Intron lengths are not shown to scale to allow alignment of exons. Only partial sequences available for *sj01* and *sj08* on BAC 18f22.

While the majority of corresponding introns are well conserved in length across the different genes on BAC 18f22, a few are highly variable (introns 4, 8, 9 and 11) (Figure 3.5). In particular the size of intron 4 was very variable, ranging from 301-13,388 bp. The fourth intron for *sj02* is 13,388 bp which is almost twice the length of any other intron. Overall, introns from the annotated BAC ranged from 49 bp to 13,388 bp with an average intron size of 710 bp. This is larger than reported for *C. elegans*, 467 bp, (Deutsch and Long, 1999) and *B. malayi*, 311 bp, (Ghedini *et al.*, 2007), consistent with the greater size of the *H. contortus* genome.

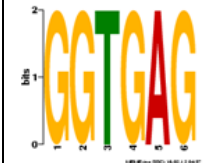

The BAC sequence was examined to determine whether repeat elements may contribute to the larger size of *H. contortus* introns. Callaghan and Beh (1994)

identified a repeat DNA sequence in *H. contortus* that was unique to this parasite. This sequence (M84609) is 150 bp in length. Hoekstra *et al.*, (1997) carried out a study to characterise *H. contortus* using microsatellites and identified a short repeat sequence termed HcREP1 (U86701). Laing *et al.*, (2011) identified sequence related to both repeat sequences tandemly arranged on *H. contortus* BAC BH4E20, and preceded by a GTCT repeat. Analysis of BAC 18f22 identified repeat sequence in the ninth intron of *sj04*. This intron is 2,634 bp in length, which is over 2 kb longer than any other ninth intron on the BAC. The Tandem Repeats Finder webpage (<http://tandem.bu.edu/trf/trf.html>) identified an upstream (GTCT)<sub>157</sub> repeat sequence 628 bp in length, significantly longer than the 56 bp GTCT repeat identified by Laing *et al.*, (2011). Five repeats with high sequence identity to the *H. contortus* repeat sequence identified by Callaghan and Beh (1994), were present 571 bp downstream from the end of the GTCT repeat (Figure 3.6). No obvious repeat sequences could be identified in other large introns.



To examine whether there was conservation of intronic sequences, MEME suite (<http://meme.sdsc.edu/meme/cgi-bin/meme.cgi>) was used. This is a motif based sequence analysis tool. Results are displayed as sequence LOGOS, which

depict the motifs as stacks of letters. The higher the height of each letter the higher the probability that a letter will appear at that position. MEME was used to compare the sequences of corresponding introns across the genes. As shown in Table 3.2, two motifs were identified, GGTGAG in intron 5 and GTAGAC in intron 7, which are highly conserved in four and three of the BAC genes respectively. Although interesting, whether these have any functional significance requires further analysis.

| Motif  | Gene and intron number                                       |
|--|--|
|   | sj065 GGTGAG<br>sj055 GGTGAG<br>sj045 GGTGAG<br>sj035 GGTGAG |
|  | sj077 CTAGAC<br>sj057 CTAGAC<br>sj037 CTAGAC                 |

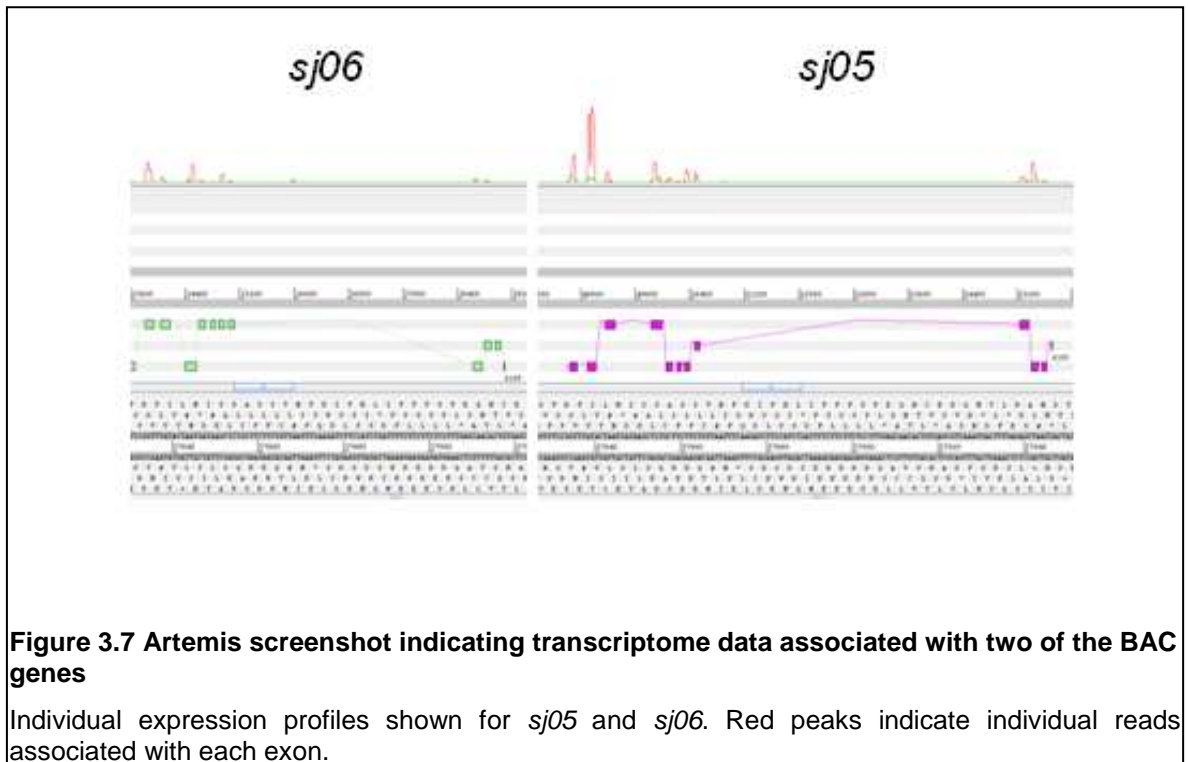
**Table 3.2 Sequence LOGOS and conserved motifs within intronic regions of the BAC genes**

Sequence LOGOS depicting the probability of any letter being present at a given site on the left, and the sequence present in each individual gene on the right, e.g. *sj065* is intron 5 in *sj06* (<http://meme.sdsc.edu/meme/cgi-bin/meme.cgi>).

#### 3.2.1.4 Expression patterns of the BAC genes

To examine the expression profiles of the protease genes encoded by the BAC, a number of approaches were taken: analysis of transcriptome data, RT-PCR and Expressed Sequence Tag (EST) analysis. Initial expression levels were identified from transcriptome data from adult worms. This data was obtained from the Sanger Institute (R. Laing, M. Berriman and J.S. Gilleard, unpublished data) and the expression profiles observed were consistent with the BAC annotation. Only transcripts associated with those regions with cysteine protease annotation were identified, thus indicating that no other genes were present on this 114 kb BAC. Additionally, a tBLASTn search of the complete *H. contortus* BAC 18f22 against the NCBI BLAST (<http://www.ncbi.nlm.nih.gov/>) did not identify any other related sequences, again confirming that there are no other genes present on the BAC. Visual observations of the transcriptome reads using Artemis indicated

that all of the genes except for *sj08* have associated transcriptome data. Subsequent analysis of L3 transcript data also failed to identify reads for *sj08*. This along with the failure to obtain a 5' RACE product for *sj08* and the lack of a stop codon in the sequence would indicate that *sj08* is a pseudogene. The expression profiles for *sj05* and *sj06* from adult stage worms are indicated in Figure 3.7.

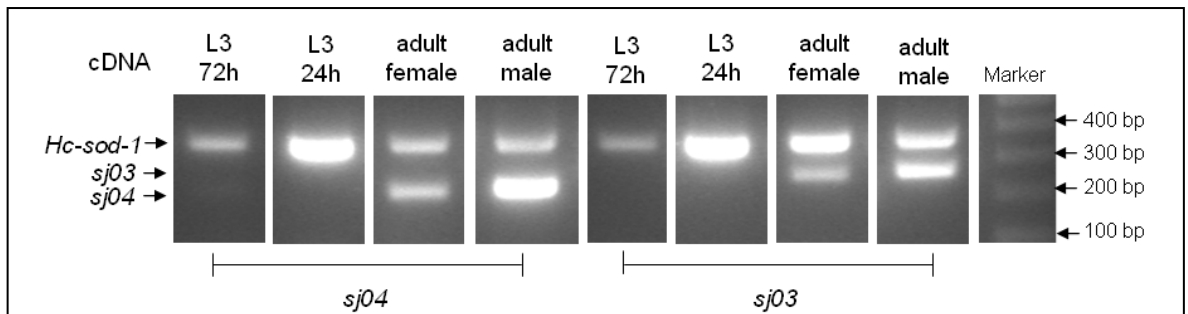


**Figure 3.7 Artemis screenshot indicating transcriptome data associated with two of the BAC genes**

Individual expression profiles shown for *sj05* and *sj06*. Red peaks indicate individual reads associated with each exon.

Expression profiles were examined by semi-quantitative RT-PCR for protease genes *sj02-sj07* on the BAC. Both *H. contortus* adult and larval stages were examined and compared relative to *Hc-sod-1* expression, which has previously been reported to be constitutively expressed in *H. contortus* (Liddell and Knox, 1998). cDNA from adult male, adult female, early L3 larvae (ex-sheathed and cultured for 24 hours) and late L3 larvae (ex-sheathed and cultured for 72 hours) (chapter 2.2.1 and 2.3.1.3) was used for PCR analysis (primers in Appendix 2 Table 3.4). The results for two of the genes, *sj03* (expected product size of 249 bp) and *sj04* (expected product size of 221 bp), are indicated in Figure 3.8. There was no expression detected in larval stages for any of the genes. This is consistent with work previously carried out on *H. contortus* CBLs for which only very low mRNA transcripts were recorded for mixed stage larvae (Pratt *et al.*, 1990).





**Figure 3.8 RT-PCR of BAC protease genes relative to *Hc-sod-1***

PCR was carried out on cDNA from different stages of *H. contortus* (chapter 2.2.1). Both gene specific and *Hc-sod-1* primers were used in each reaction and products were separated by 2% agarose gel electrophoresis and stained with 0.05 µg/ml ethidium bromide. PCR was carried out for BAC genes *sj02-sj07* however repeatable results observed only for *sj03* and *sj04*.

There is an abundance of Expressed Sequence Tag (EST) data available for a number of organisms at varying lifecycle stages. ESTs are obtained from cDNA sequence reads and indicate expressed genes. As only one read is obtained for each region of the genome there are often errors in the sequence (Parkinson *et al.*, 2002). The Nembase webpage (<http://www.nematodes.org/nembase4/>) has 22,257 identified ESTs in its database of *H. contortus* adult worms, and these have been grouped into 4,972 clusters. ESTs are clustered to reduce redundancy creating one transcript per gene (Nagaraj *et al.*, 2007). Expression data was also studied using the information available on the Washington University Basic Local Alignment Tool (WU-BLAST) ([www.ebi.ac.uk/Tools/sss/wublast/parasites.html](http://www.ebi.ac.uk/Tools/sss/wublast/parasites.html)). BLAST searches were carried out against both databases taking the full cDNA sequence of each of the genes. A 95% sequence identity was used as a cut off in identification of potential matches. The BLAST searches carried out against Wash U identified sequences highly similar to *sj03*, *sj04*, *sj05* and *sj06* using this criterion and similarly *sj03*, *sj04* and *sj05* were identified using Nembase 4 (Table 3.3). If the sequence identity was reduced to 94%, sequence representing *sj06* could be identified in Nembase.

Geldof *et al.*, (2005) identified that the most abundantly represented genes in the Nembase EST dataset are *Hc-nim-1* and *Hc-nim-2*, with these having the highest number of associated ESTs. Out of the 39 clusters that were identified as being the most abundantly expressed in the *H. contortus* dataset, one of these clusters, HCC00016 with 72 associated ESTs, has high identity to *sj04*, indicating that this is the most abundantly expressed cysteine protease gene present on the

BAC. Therefore transcriptome, RT-PCR and EST analysis indicate that, of the BAC genes, *sj04* is the most highly expressed, then *sj05*, *sj03*, *sj06*, *sj07* and *sj02*, with initial transcriptome observations being supported by EST data.

| Gene        | Nembase cluster number | Percentage DNA identity | Nembase coverage/gene length (bp) | Number of ESTs |
|-------------|------------------------|-------------------------|-----------------------------------|----------------|
| <i>sj01</i> | -                      | -                       | -                                 | -              |
| <i>sj02</i> | -                      | -                       | -                                 | -              |
| <i>sj03</i> | HCC01618_1             | 95                      | 727/1023                          | 7              |
| <i>sj04</i> | HCC00016_2             | 97                      | 100/1026                          | 72             |
| <i>sj05</i> | HCC00024_1             | 95                      | 972/1023                          | 7              |
| <i>sj06</i> | HCC00562_1             | 94                      | 524/1029                          | 6              |
| <i>sj07</i> | -                      | -                       | -                                 | -              |
| <i>sj08</i> | -                      | -                       | -                                 | -              |

**Table 3.3 EST data for the BAC genes obtained from Nembase 4**

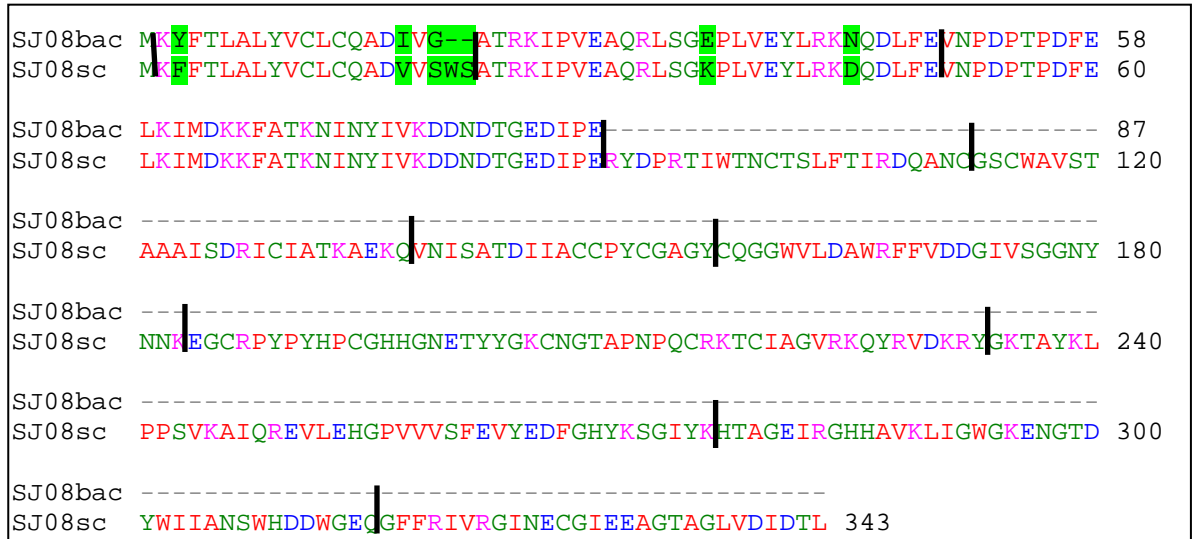
Information obtained from Nembase BLAST searches using the maximum sequence information available for each gene. All cluster sequences from adult *H. contortus* stage ([www.nematodes.org/nembase4/](http://www.nematodes.org/nembase4/)).

### 3.2.1.5 Identification of a complete *sj08* sequence

Previously, the complete sequence for the last gene on the BAC (*sj08*) was not available due to its location near the end of the BAC. A tBLASTn search using the amino acid sequence of the three exons from *sj08* that were present on the BAC against the database of *H. contortus* assembled supercontigs (21/08/08)(all reads) was carried out. Two supercontigs were identified as having high sequence identity to *sj08*. Supercontig\_0047752 had the higher percentage conservation due to its similarity to the first three exons present on the BAC, there were however no additional regions on this supercontig that gave any indication that a cysteine protease was present. Subsequent study of supercontig\_0058857 indicated that the complete *sj08* gene was present on this supercontig. There are a number of bp differences between the sequence on the BAC and this supercontig, however there is high enough conservation to suggest that this is the same gene (92% DNA identity in the first 3 exons) (Figure 3.9). Further searches using the BAC gene sequences and active cysteine protease regions against this supercontig gave no other hits, suggesting that there are no other related genes downstream of *sj08* on this supercontig. Several attempts to confirm linkage of the BAC and supercontig sequence by PCR failed to produce a

product. It is possible this could be due to variation in the region used for primer design and/or low gene expression or a pseudogene.

In addition to a BLAST search for the end of *sj08*, a BLAST search was performed to identify the start of *sj01*. No overlapping supercontigs were identified and *sj01* is currently incomplete.



**Figure 3.9 SJ08 sequence alignments for the BAC 18f22 and supercontig\_0058857**

Alignment includes the first three exons present on the BAC (SJ08bac) and the full protease sequence encoded by supercontig\_0058857 (SJ08sc). Lines indicate the intron/exon boundaries, green shading indicates bp changes. Amino acid colours reflect physicochemical properties; red for small, blue for acidic, magenta for basic and green for hydroxyl + sulfhydryl + amine + G.

## 3.2.2 Analysis of proteases encoded by *H. contortus* BAC 18f22

### 3.2.2.1 Similarity of BAC-encoded proteins to the previously identified *H. contortus* AC family

The eight cysteine proteases encoded by BAC 18f22 were identified by tBLASTn analysis using the AC-2 protease sequence, previously identified by Pratt *et al.*, (1990). To examine the similarity of proteases encoded by the BAC to one another and to the AC family of which there are five members (AC-1-AC-5), the percentage identity and similarity across the proteins was calculated and is shown in Table 3.4. This table indicates a 97% sequence conservation between AC-1 and AC-2. Additionally, there is a 97% identity between SJ04 & AC-1 and a 98% identity between SJ04 & AC-2. Analysis at the DNA level indicated a 98%

identity between *AC-1* and *sj04*, thus it is proposed that that *sj04*, *AC-1* and *AC-2* are the same gene. Another high percentage DNA identity, 94%, is noted between *SJ05* and *AC-3*, which also suggests that these genes are the same. There were no other instances of very high conservation between encoded proteases. Therefore it is likely that all others represent novel and distinct proteins. Thus, given that *AC-1*, *AC-2* & *SJ04*, and *AC-3* & *SJ05* appear to be encoded by the same genes, this family contains at least 10 members.

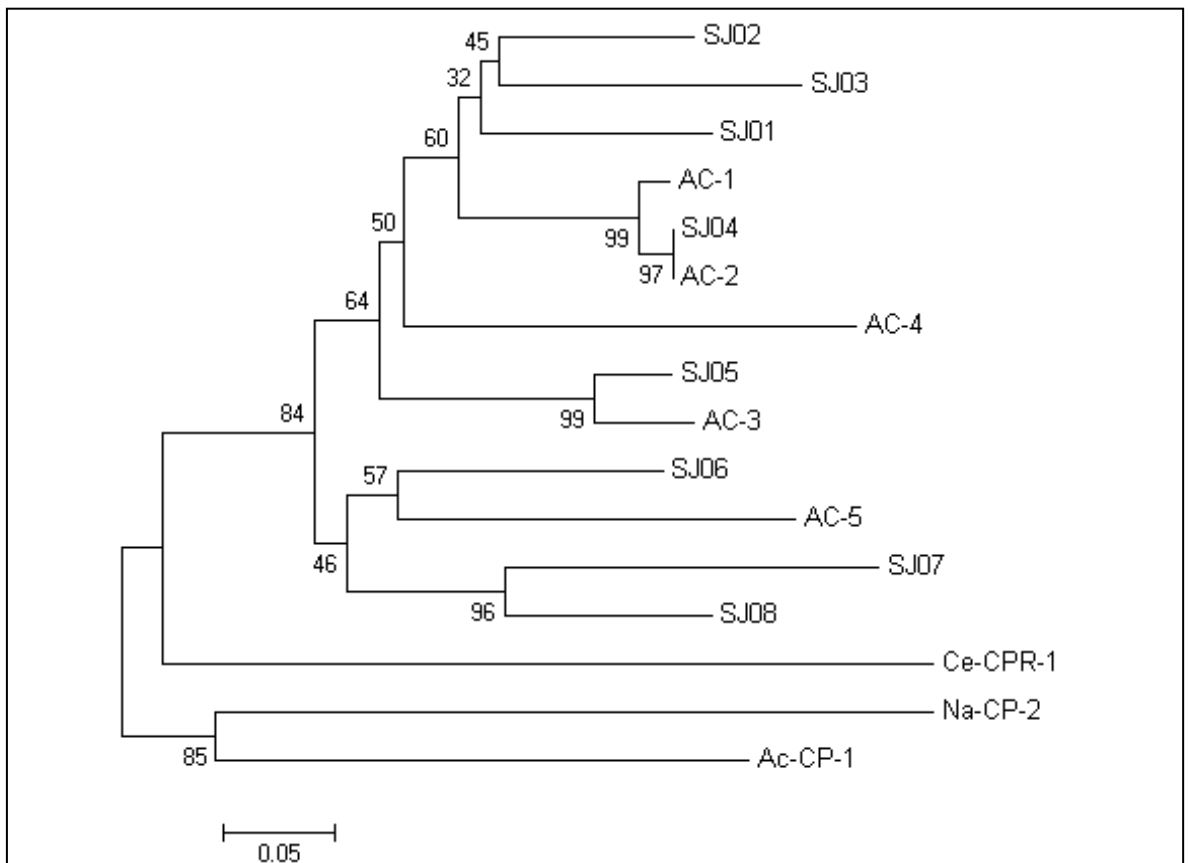
Analysis of the percentage identities between different family members showed that in some cases those positioned consecutively on the BAC were more similar to one another (Table 3.4). This is true of proteases *SJ02* & *SJ03*, *SJ04* & *SJ05*, *SJ06* & *SJ07* and *SJ07* & *SJ08*, consistent with duplication and sequence divergence within this gene region.

| Protease | AC-1 | AC-2  | AC-3  | AC-4  | AC-5  | SJ02  | SJ03  | SJ04  | SJ05  | SJ06  | SJ07  | SJ08  |
|----------|------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| AC-1     | 100  | 97/98 | 75/85 | 75/84 | 65/77 | 78/87 | 73/85 | 97/98 | 75/85 | 68/81 | 66/79 | 65/78 |
| AC-2     |      | 100   | 75/85 | 74/84 | 64/76 | 78/87 | 73/85 | 98/98 | 75/85 | 69/81 | 66/79 | 66/79 |
| AC-3     |      |       | 100   | 73/83 | 62/73 | 71/80 | 74/83 | 72/82 | 91/93 | 68/78 | 63/75 | 64/75 |
| AC-4     |      |       |       | 100   | 61/72 | 71/82 | 71/82 | 72/82 | 72/82 | 62/75 | 60/73 | 59/72 |
| AC-5     |      |       |       |       | 100   | 59/71 | 59/72 | 62/74 | 62/73 | 63/72 | 64/75 | 59/71 |
| SJ02     |      |       |       |       |       | 100   | 71/81 | 79/88 | 74/84 | 69/80 | 67/78 | 65/78 |
| SJ03     |      |       |       |       |       |       | 100   | 74/86 | 76/87 | 65/79 | 65/76 | 65/77 |
| SJ04     |      |       |       |       |       |       |       | 100   | 75/85 | 69/81 | 66/79 | 66/79 |
| SJ05     |      |       |       |       |       |       |       |       | 100   | 71/82 | 66/79 | 66/78 |
| SJ06     |      |       |       |       |       |       |       |       |       | 100   | 70/81 | 66/79 |
| SJ07     |      |       |       |       |       |       |       |       |       |       | 100   | 76/86 |
| SJ08     |      |       |       |       |       |       |       |       |       |       |       | 100   |

**Table 3.4 Percentage identity/similarity at the amino acid level of proteases encoded by *H. contortus* BAC 18f22 and the AC protease family**

Percentages highlighted in red are those indicating high amino acid identity and similarity between protein sequences.

A similar pattern was depicted using the Molecular Evolutionary Genetic Analysis (MEGA) tool (version 5) which is available to download at <http://www.megasoftware.net/>, indicating that proteases encoded by genes that are located closer together on the BAC, group closer together phylogenetically. This is evident in Figure 3.10, with the proteases encoded by the BAC spanning the length of the tree in almost the exact order in which they are found on the BAC. This observation is interesting as both AC-4 and AC-5 are located on the tree within the region covered by the proteases encoded by the BAC and as the rest of the proteases appear in almost perfect chronological order this suggests that AC-4 and AC-5 should be located on the BAC. Additionally, as expected, the proteases thought to be represented in duplicate group together; AC-1, AC-2 & SJ04 and AC-3 & SJ05. Figure 3.10 is a phylogram and contains additional information about the divergence of proteases, with those furthest away from the node being the most divergent.



**Figure 3.10 Phylogenetic tree indicating the BAC and AC protease families**

This is a MEGA phylogram, in which branch length is proportional to divergence. A scale bar is included to indicate the degree of change. Bootstrapping has been applied to the tree to confirm reliability. Additional proteases are shown from *C. elegans*, *Necator americanus* and *Ancylostoma caninum*.

A tBLASTn search using the published amino acid sequence for AC-4 and AC-5, was carried out against the database of *H. contortus* supercontigs (28/08/09). Sequence encoding AC-4 was identified on supercontig\_0005737 however unsequenced regions prevented identification of the full protease. Only 244 out of the full 343 amino acids were identified, the missing region could not be identified in any part of the genome and thus it is likely to be within an unsequenced region. This partial sequence is very highly conserved with AC-4 (Pratt *et al.*, 1992), with only 3 amino acid changes in the 244 identified amino acids (Figure 3.11), and a 98% sequence identity at the DNA level. AC-5 could be fully annotated on supercontig\_0008756. Comparing the published amino acid sequence of AC-5 (Pratt *et al.*, 1992) and the AC-5 sequence identified from the supercontig indicated that there are 25 amino acid differences between the two (Figure 3.12), and a 94% sequence identity at the DNA level. This indicates that both proteases are present in UK isolates, but as yet their genomic locations have not been linked to the other proteases within the family.

|        |   |     |
|--------|---|-----|
| AC-4sc | MYLVLTLCAYLCAASGASINAAQEIPLEAQTTLTGEPLVAYLRKNQNLFEVNSEPTPNYEQ | 60  |
| AC-4   | MYLVLTLCAYLCAASGASINAAQEIPLEAQTTLTGEPLVAYLRKNQNLFEVNSEPTPNYEQ | 60  |
| AC-4sc | KIMDIKFKNQKLNQVVKNDPEPNEDIPEEYDPREKFKCSTFYIRDQANCGSCWAVSTAA   | 119 |
| AC-4   | KIMDIKFKNQKLNQVVKNDPEPNEDIPEEYDPREKFKCSTFYIRDQANCGSCWAVSTAA   | 120 |
| AC-4sc | AISDRICIATNGE-----  | 132 |
| AC-4   | AISDRICIATNGEKQVNISSTDILTCCNPQCGFGCGGWSIRAWEYFVYEGVVS GGGEYLT | 180 |
| AC-4sc | -----GKVAYGVPEP   | 141 |
| AC-4   | KGVCRPYPPIHPCGHHGNDTYYGECPREAATPPCKKKCQPGYKKIFRMDKRQGVAYGVPEP | 240 |
| AC-4sc | KEEAIQREILRHGPVVASFVYEDFSLYKTGVYKHTAGALRGYHAVKMMGWGVDSKTKAK   | 201 |
| AC-4   | KEEAIQREILRHGPVVASFVYEDFSLYKTGVYKHTAGALRGYHAVKMMGWGVDSKTKAK   | 300 |
| AC-4sc | YWLIANSWHNDWGENGYFRFIRGINDCEIEDTVAAGIVDVDSL                   | 244 |
| AC-4   | YWLIANSWHNDWGENGYFRFIRGINDCEIEDTVAAGIVDVDSL                   | 343 |

**Figure 3.11 AC-4 sequence comparison with supercontig\_0005737**

Amino acid alignment of the previously identified AC-4 protease and the protease identified on supercontig\_0005737. Amino acid changes highlighted in green.

|        |  |     |
|--------|--|-----|
| AC-5sc | MRYLVLALYLYLCRTLGLADTDLAQGIPLHAQMLTGAPLVEYLQKNQDLFEVRRTPPTPGFK | 60  |
| AC-5   | MRHIVLALYLYLCRTLGLADTDLAQGIPLHAQMLTGAPLVEYLQKNQDLFEVRRTPPTPGFK | 60  |
| AC-5sc | YKLMDKAFANANQNLPVVNDNDNTGADLPESYDPRIVWENCSSFHIRDQANCGSCWAV     | 120 |
| AC-5   | YKLMDKAFANANQNLPVVNDNDNTGADLPENYDPRIVWENCSSFHIRDQANCGSCWAV     | 120 |
| AC-5sc | STAAAIHDRICCIATKGGKQVYASDITDILTCCGAPCGMGCRGGWPIEAWKFFEYDGVVSGG | 180 |
| AC-5   | STAAAIHDRICCIATKGGKQVYASDITDILTCCGAPCGMGCRGGWPIEAWKFFEYDGVVSGG | 180 |
| AC-5sc | PYLGKGCCSPYPLHPCGRHGNDTFYGNCAGMAATPPCKRCQPGFRGMYRVDKRYGESRK    | 240 |
| AC-5   | PYLGKGCCSPYPLHPCGRHGNDTFYGNCAVGMATPPCKRCQPGFRGMYRVDKRYGEPGR    | 240 |
| AC-5sc | AYRLPSSEVKIRRDIMERGSVVAVFAVYEDFSHYQSGIYKHTAGRFTGGYHAVKMIGWGK   | 300 |
| AC-5   | TYTLPRSEVKIRRDIMERGSVVAVFAVYEDFSHYQSGIYKHTAGRFTGGYHAVKMIGWGK   | 300 |
| AC-5sc | DNGTDYWLIANSWHDDWGNGFFRMIRGINNCGIEEQVDAGIVDVESL                | 348 |
| AC-5   | DNGTDYWLIANSWHDDWGNGFFRMIRGINNCGIEEQVDAGIVDVESL                | 348 |

**Figure 3.12 AC-5 sequence comparison with supercontig\_0008756**

Amino acid alignment of the previously identified AC-5 protease and the protease identified on supercontig\_0008756. Amino acid changes highlighted in green.

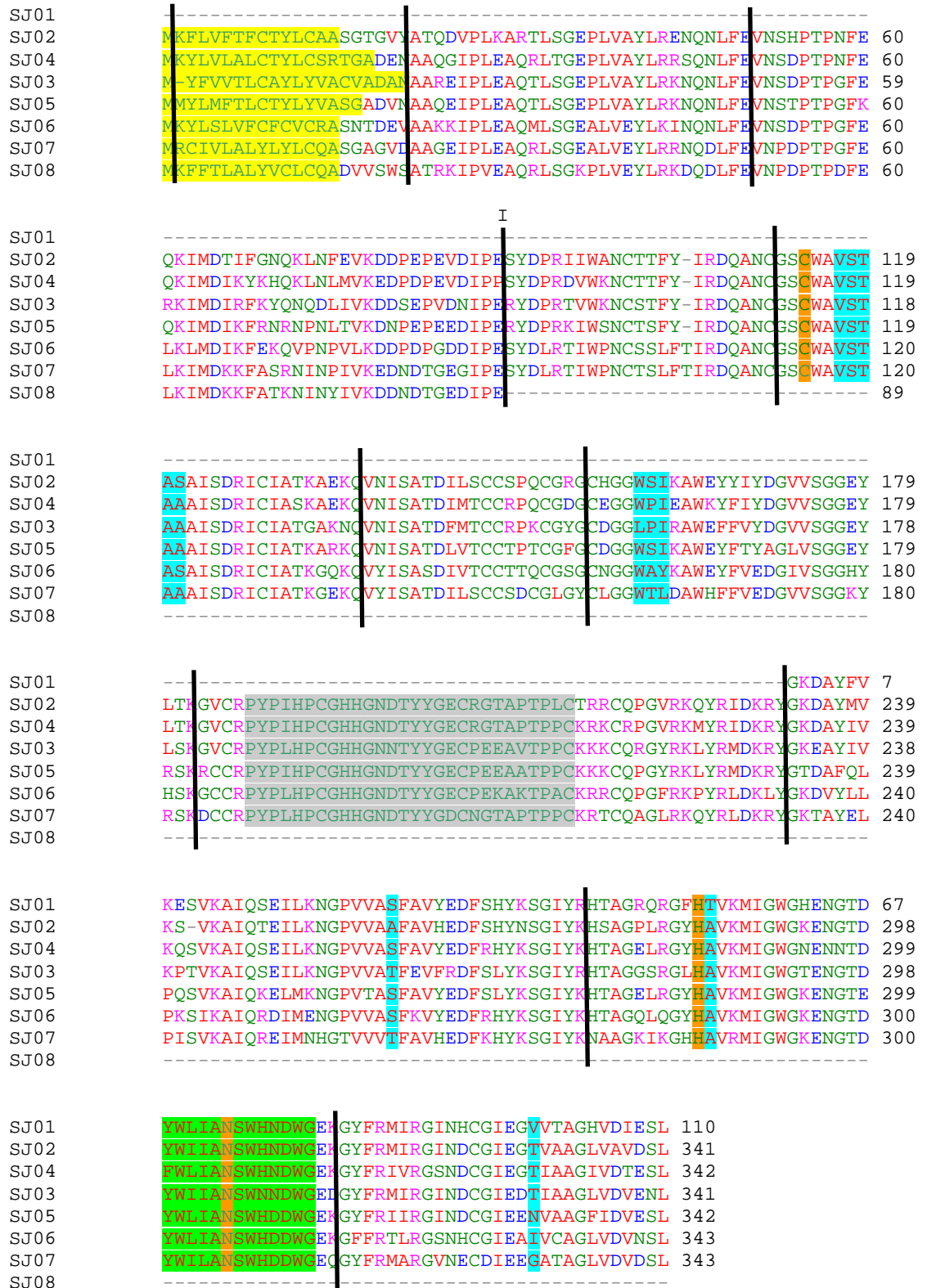
### 3.2.2.2 Structure and function of cysteine proteases

Cysteine proteases contain a number of important conserved amino acid domains within their structures. Sequence comparisons were carried out and the sequences aligned using Clustal W2 (<http://www.ebi.ac.uk/Tools/msa/clustalw2/>) (Figure 3.13). There are three active site residues; cysteine, histidine and asparagine, which are essential for activity. Rehman and Jasmer (1999) identified a number of structural features relating to the cysteinyl active sites that can be used for classification within the cysteine protease family. This work identified that in Cathepsin B proteases, the cysteinyl active site is followed by one of two predominant signature sequences (Type A; FGAVE or Type B; VSTAA). For the proteases studied, Type A was found to be most common in *C. elegans* and Type B in *H. contortus*. The BAC-encoded and AC proteases are all Type B, with only a substitution of A to S in SJ02 and SJ06. CBL cysteine proteases also contain a 28 AA ‘occluding loop’ which is responsible for blocking off one end of the substrate binding cleft, contributing to enzyme stability and conferring exopeptidase activity (Illy *et al.*, 1997). Three S2 subsites in cysteine proteases are also important for enzyme specificity, with these regions interacting with substrate residues P<sub>1</sub>-P<sub>3</sub> (Illy *et al.*, 1997). In a study carried out by Rehman and Jasmer (1999) using rat cathepsin B sequence (P00787) as a reference, the identified subsites were; tyr75, pro76, ser77, ala173, ala200 and glu245 (or Y, P, S, A, A and E in the rat cathepsin B



sequence), with numbering based on the mature enzyme. The Y, P and S are located next to one another and the A, A and E at separate locations. As can be observed from Figure 3.13 and Table 3.5, the cysteinyl active site signature sequence for all of the BAC-encoded and AC proteases is very highly conserved, with only the one amino acid change in SJ02 and SJ06. Although some of the S2 sites are conserved, there are also a number of amino acid changes, particularly at the YPS (75-77) and E245 subsites, which may have consequences for substrate specificity.

In addition to the conserved active and S2 subsites, there is a proposed haemoglobinase motif (YWLVANSW--DWGD) located around the region of the asparagine active site (294-307). It has been suggested that enzymes containing this motif are able to degrade haemoglobin based on their expression in blood feeding stages (Baig *et al.*, 2006). This haemoglobinase motif is highly conserved in the CBLs of helminth blood-feeders, though AC-1 and AC-2 (and therefore SJ04) have a variation at position 1 of this motif. The substitution of Y to F is speculated to alter the function of the motif and these proteins may provide more of a housekeeping role (Baig *et al.*, 2006).



**Figure 3.13** Amino acid alignment of the *H. contortus* proteases encoded by BAC 18f22

The signal peptide region is highlighted in yellow, the putative pro-region cleavage site indicated by I, the active site regions in orange and the S2 subsites in blue. The occluding loop region is highlighted in grey and the haemoglobinase motif indicated in green. Intron/exon boundaries are indicated by the black lines and un-sequenced regions, in SJ01 and SJ08, by dashed lines.

| Sequence | Cysteiny active site signature sequence (111-115) | S2 Subsite              |         |         |         | Haemoglobinase motif (301-315) |
|----------|---|-------------------------|---------|---------|---------|--------------------------------|
|          |   | Y, P, S (160, 161, 162) | A (258) | A (285) | E (332) |                                |
| ratcatB  | FGAVE   | YPS                     | A       | A       | E       | YWLVANSW--DWGD                 |
| SJ01     | -   | -                       | S       | T       | V       | YWLIANSW--DWGE                 |
| SJ02     | VSTAS   | WSI                     | A       | A       | T       | YWIIANSW--DWGE                 |
| SJ03     | VSTAA   | LPI                     | T       | A       | T       | YWIIANSW--DWGE                 |
| SJ04     | VSTAA   | WPI                     | S       | A       | T       | FWLIANSW--DWGE                 |
| SJ05     | VSTAA   | WSI                     | S       | A       | N       | YWLIANSW--DWGE                 |
| SJ06     | VSTAS   | WAY                     | S       | A       | I       | YWLIANSW--DWGE                 |
| SJ07     | VSTAA   | WTL                     | T       | A       | G       | YWILANSW--DWGE                 |
| SJ08     | VSTAA   | WVL                     | V       | A       | A       | YWIIANSW--DWGE                 |
| AC-1     | VSTAA   | WPI                     | S       | A       | T       | FWLIANSW--DWGE                 |
| AC-2     | VSTAA   | WPI                     | S       | A       | T       | FWLIANSW--DWGE                 |
| AC-3     | VSTAA   | WPI                     | S       | A       | N       | YWLIANSW--DWGE                 |
| AC-4     | VSTAA   | WPI                     | S       | A       | T       | YWLIANSW--DWGE                 |
| AC-5     | VSTAA   | WPI                     | V       | A       | Q       | YWLIANSW--DWGE                 |

**Table 3.5 Conserved regions within the BAC and AC CBL cysteine protease families**

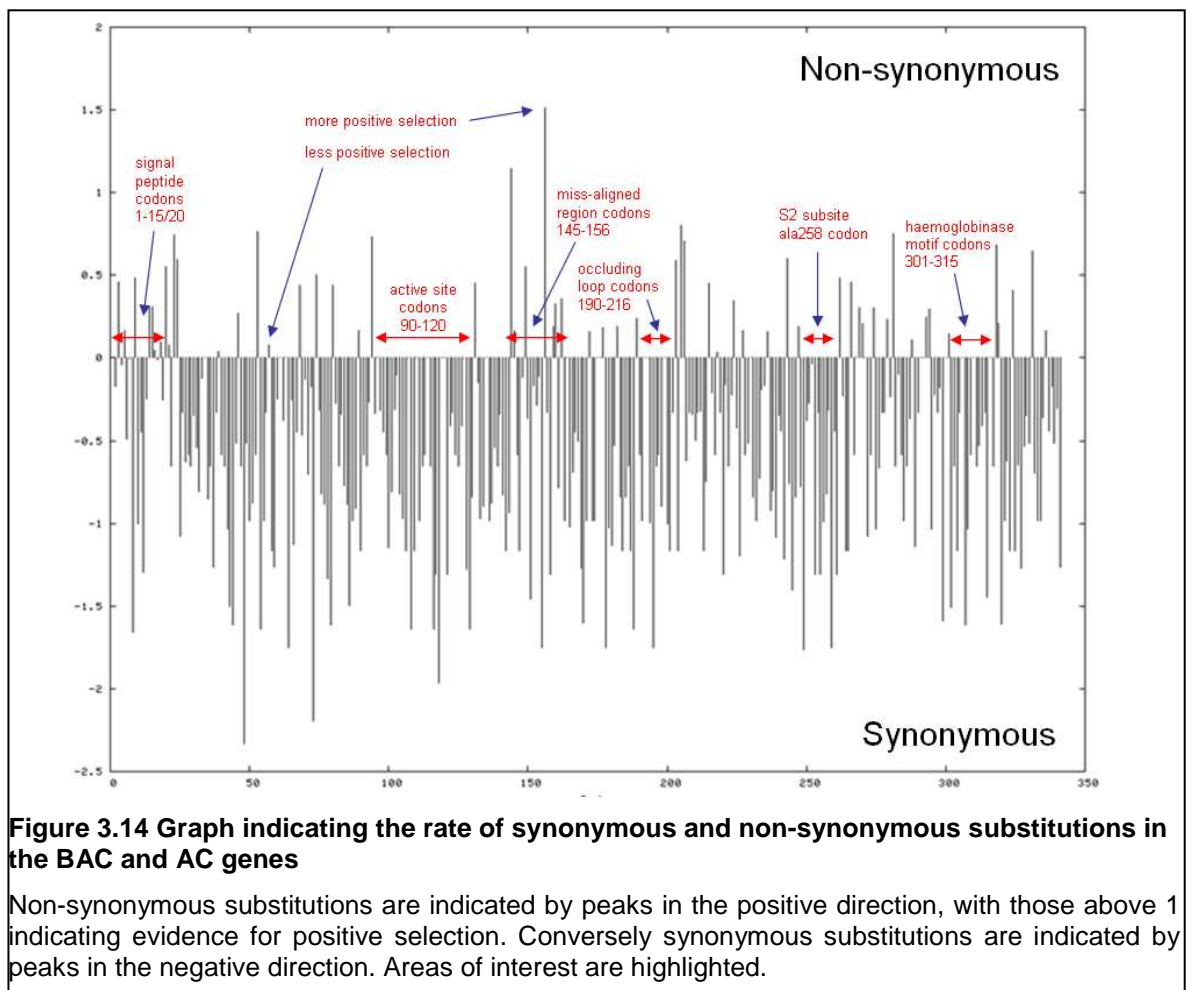
Regions in the haemoglobinase motif in which tyr294 (Y) have been replaced with phe294 (F) have been highlighted in bold.

### 3.2.2.3 Analysis of synonymous and non-synonymous substitutions in cathepsin B cysteine proteases

It was of interest to determine if any regions of the proteases showed any evidence of selection for amino acid diversity which could potentially alter function and/or antigenicity. Datamonkey (<http://www.datamonkey.org/dataupload.php>) is used to detect positive selection and the rate of synonymous (silent) and non-synonymous (where nucleotide changes alter the amino acid) substitutions in genes. For this analysis the full BAC exonic gene sequences of *sj02-sj08* were used. All input sequences are required to be of the same length to allow the programme to function, and for this reason the first codon of exon 2 was removed in genes *sj02*, *sj04*, *sj05*, *sj06*, *sj07* and *sj08* as they all contain one codon more than *sj03*. For this reason one codon in exon 5 of *sj06*, *sj07* and *sj08* was also removed. Figure 3.14 obtained from this analysis indicates regions that are highly variable if there is a peak in the positive direction and conserved if the peak is in the negative direction. Peaks in the positive direction occur if the rate of non-synonymous substitution is higher than synonymous. Non-synonymous substitutions are more likely to occur by chance, and have an effect on the function of the protein, compared to synonymous substitutions. The higher the peaks in the positive direction the more evidence there is for positive selection in that region and subsequent advantageous genetic diversity. From analysis of the amino acid alignments of the BAC encoded and AC proteases it is easy to identify regions that are less well conserved, however visual analysis of the DNA sequence to detect base pair changes resulting in non-synonymous substitutions is more difficult. The Datamonkey programme is useful for detecting synonymous substitutions that are not apparent on analysis of the amino acid sequence alone; these may or may not be significant.

Figure 3.14 indicates the rate of synonymous and non-synonymous substitutions in the BAC genes *sj02-sj08* and the AC genes *AC-1*, *AC-3*, *AC-4* and *AC-5*. Although some of these genes were duplicated, for example *AC-1* & *sj04* and *AC-3* & *sj05*, they were added to the analysis to help increase the reliability of the results. Ideally at least 10 different sequences would be used to increase the reliability of the results. This was not possible as there are a limited number of genes on BAC 18f22 and in the AC family that are not duplicated. Areas of

interest have been added to this figure to aid interpretation. Within the signal peptide region there are a number of peaks in the positive direction indicating variation in this region. This finding is not surprising as the signal peptide, which is cleaved and therefore not involved in protease function, is commonly a very variable region. The size of the peaks in the positive direction is also indicative of the degree of positive selection, with the higher peaks having more positive selection. In addition to this, the region from codons 145-156 has a number of peaks in the positive direction, indicating variation. In alignment of the BAC genes this region appears to misalign and leave gaps in the sequence in order to maintain the highest percentage identity. It is therefore unsurprising that this region is highlighted as having numerous non-synonymous substitutions. Other areas of interest include the 'GSCWAV' active site, the occluding loop region, S2 subsites and the region of the haemoglobinase motif, where there are no peaks indicating non-synonymous substitutions, but are peaks in the negative direction indicative of synonymous substitutions.



### 3.2.3 Naming the genes encoded by BAC 18f22

Work presented within this chapter indicates that the previously characterised AC gene family (Pratt *et al.*, 1992) has additional members identified on *H. contortus* BAC 18f22. Evidence suggests that AC-1, AC-2 & SJ04 are encoded by the same gene and AC-3 & SJ05 are encoded by the same gene and thus are already published sequences. Table 3.6 indicates the BAC names for which the genes will be known, indicating those which have high sequence identity to already published sequences.

| Initial gene name used (this work) | New BAC gene name    | Associated published sequence |
|------------------------------------|----------------------|-------------------------------|
| <i>sj01</i>                        | <i>hc-BAC18f22-1</i> | -                             |
| <i>sj02</i>                        | <i>hc-BAC18f22-2</i> | -                             |
| <i>sj03</i>                        | <i>hc-BAC18f22-3</i> | -                             |
| <i>sj04</i>                        | <i>hc-BAC18f22-4</i> | AC-1, AC-2                    |
| <i>sj05</i>                        | <i>hc-BAC18f22-5</i> | AC-3                          |
| <i>sj06</i>                        | <i>hc-BAC18f22-6</i> | -                             |
| <i>sj07</i>                        | <i>hc-BAC18f22-7</i> | -                             |
| <i>sj08</i>                        | <i>hc-BAC18f22-8</i> | -                             |

**Table 3.6 Re-naming the genes present on *H. contortus* BAC18f22**

### 3.3 Discussion

The aim of this chapter was to use the available BAC sequence information to identify and characterise a *H. contortus* cathepsin B-like (CBL) protease gene family. Six novel *H. contortus* CBL cysteine protease genes were identified, and both bioinformatic and phylogenetic analysis indicate that they are additional members of the AC protease gene family. RT-PCR analysis of the BAC protease genes identified expression in adult but not infective larvae stages, a finding which is consistent with previous analysis of the AC protease gene family (Pratt *et al.*, 1990). This could be due to lifecycle adaptation, as transcripts for these proteases are enriched in the intestine (Jasmer *et al.*, 2001) and are likely to have a role in digestion. For this reason their requirement in the young larval/non-blood feeding stages would be redundant (Ray and McKerrow, 1992).

CBLs have received considerable attention as vaccine candidates or drug targets due to their abundance and speculated role in blood feeding (Pratt *et al.*, 1990). Work carried out by Jasmer *et al.*, (2001) identified that 17% of *H. contortus* intestinal transcripts (from EST data at that time) were represented by CBL genes. De Vries *et al.*, (2009) carried out a vaccine trial using excretory/secretory (ES) products from *H. contortus* enriched with the AC-5 CBL cysteine protease, as previous work indicated protective properties of ES antigens enriched for cysteine proteases (Bakker *et al.*, 2004). A reduction in egg output was identified when using AC-5 enriched products, and together with previous evidence, indicates the potential protective effect afforded by vaccination with cysteine proteases.

Work has been carried out looking at CBL cysteine proteases in other blood feeding parasites, namely the hookworms *A. caninum* and *N. americanus* (Ranjit *et al.*, 2008). Ranjit *et al.*, (2008) identified a family of four CBL cysteine proteases in *N. americanus*, Na-CP-2, -3, -4, -5, which are expressed in the intestine and considered potential vaccine targets. Phylogenetic analysis indicated that this family is related to CBLs in other nematodes, most closely to *A. caninum* (Ac-CP-1 and Ac-CP-2) and the *H. contortus* HmCP protease family (Skuce *et al.*, 1999). Loukas *et al.*, (2004) examined the effect of vaccination of dogs with Ac-CP-2. Ac-CP-2 is suggested to be involved in haemoglobin digestion

and was chosen as a vaccine candidate for this reason. Expression as an active enzyme resulted in reduced fecal egg counts, a decrease in worm size and a decrease in the ratio of female to male worms. Antibodies were generated and bound to the intestinal lumen of the hookworms, thus it is expected that they cause interference with protein function. Although no decrease in worm burden was found, vaccination reduced clinical symptoms.

In this study, annotation and analysis of BAC 18f22 showed that the genes encoding AC proteases are tandemly arranged within the genome, with eight AC genes contained in a 114 kb genomic region. It is a common occurrence, across a number of parasitic and free-living nematode species, for CBL cysteine protease genes to be present as multigene families (Larminie and Johnstone, 1996; Shompole and Jasmer, 2001). However, with the exception of *Caenorhabditis* species, it was not known how these are organised in the genome. In *C. elegans*, the identified cysteine proteases are not tandemly arranged in the genome. Data presented here indicates that members of the *H. contortus* AC protease family have arisen by duplication and divergence of a common ancestor. The tandem arrangement of these genes suggests this is likely to be a recent duplication and/or there is selection for those to be in the same genomic region, possibly for co-ordinated expression.

The individual genes within multigene families may have different functions and gene family organisation may therefore indicate how diverse these functions are e.g. the globulin gene family in mammals. Members of the globulin gene family are all located in the same region of the genome, suggesting that their expression is co-ordinated. Different forms of haemoglobin result from expression of different members of the gene family (Walsh and Stephan, 2008). Pratt *et al.*, (1992) hypothesised that the occurrence of the AC protease family in *H. contortus* may be due to the requirement for a large quantity of protease in a short period of time, a requirement for functional diversity or alternatively, diversity in antigenicity to avoid immune recognition. Analysis of the amino acid residues around the active site regions, the S2 subsites and putative haemoglobinase domain, revealed some alterations. Whether these give rise to variations in activity and/or antigenicity requires biochemical and immunological studies.



5' RACE confirmed the correct location of the start codon, which had been difficult to identify bioinformatically as the first exon contained only the start methionine. This feature was first noted in the AC proteases identified by Pratt *et al.*, (1990) and has been identified in other parasitic nematodes, and in the *C. elegans* cysteine proteases CPR-1, CPR-4 and CPR6 (Larminie and Johnstone, 1996). It is unknown how widely or frequently this short first exon is found in parasitic species, if it is a feature found in all gene types and ultimately if it has any significance on protease expression.

Analysis of introns present in the BAC genes was carried out as previous work indicated that they may have important roles in gene expression. Work carried out by Nam *et al.*, (2002) identified a 7.2 kb intron in the *C. elegans run* gene that is important for control of gene expression. This gene is a homologue of the *Drosophila runt* gene. The mammalian homologues of this gene have roles in haematopoiesis and osteogenesis, and in *C. elegans run* was detected in the intestine. Deletion constructs were generated (with varying sizes of deleted intronic sequence) and showed variation in expression levels. The presence of large introns in *C. elegans* is unusual and thus there may be a functional relationship between these large introns and those observed in *H. contortus*. In addition, a number of introns have been identified as causing an intron-mediated enhancement (IME) of genes (Mascarenhas *et al.*, 1990). Rose *et al.*, (2008) explored IME in *Arabidopsis thaliana* genes, by carrying out bioinformatic analysis on a number of introns to predict their enhancing ability. A number of motifs were identified as being present in these introns and potentially having a role in IME. The *A. thaliana* intron motifs were compared to those identified in the BAC introns, however there was no consensus. This was not unexpected given the divergence between these organisms. Work carried out by Ho *et al.*, (2001) and Okkema *et al.*, (1993) suggests that *C. elegans* may also display IME effects as intron inclusion in expression work often leads to an increase in mRNA and protein. This is also found in reporter genes such as Lac-Z and GFP, where introns enhance expression (Fire *et al.*, 1990). As *H. contortus* genes have been identified as having an increased number and size of introns compared to *C. elegans*, any IME occurring in *H. contortus* may be significant. Large scale examination would need to be carried out to determine this, as data obtained to

date has not indicated any obvious relationship between intron size and expression level in the *H. contortus* BAC genes.

As a pattern, most of the intron sizes are well conserved in the AC genes, although a few introns were of varied length. The first intron was relatively short in all AC genes which is in contrast to the findings of Bradnam and Korf (2008) who indicated that in the model species *Arabidopsis thaliana*, *Drosophila melanogaster* and *C. elegans*, longer first introns are a general property. Previous work has also suggested that early introns may have important roles in the function of gene expression and for this reason are significant regions for analysis (Mascarenhas *et al.*, 1990). The fourth intron of the BAC genes, with the exception of *sj07*, is longer than the other introns (2,190-13,388 bp). This intron is located at the junction between the pro and mature enzyme and could potentially have a function in regulating processing of the mature enzyme. During intron analysis a repeat sequence was identified in intron nine of *sj04*. This repeat sequence shares >80% identity to repeat sequence M84609, identified in *H. contortus* by Callaghan and Beh (1994) and is preceded by repeats of GTCT. Hoekstra *et al.*, (1997) identified another repeat (HcREP1) that shares only a 48% sequence identity with the repeat sequence M84609. Interestingly, Laing *et al.*, (2011) identified repeats related to HcREP1 on the *H. contortus* BAC BH4E20, at a duplication break point, between two closely related genes. In the work presented here, neither HcREP1 nor M84609 repeat sequences were found between the different AC genes on the BAC, suggesting repeat sequences are not contributing to gene duplication events in this case. A BLASTn was carried out using M84609 against the *H. contortus* assembled supercontigs (21/08/08)(all reads) on the Sanger webpage to determine if this sequence is present in other parts of the genome. Over seventy supercontigs were identified as having the sequence present, it is repeated and shares over >80% identity in the majority of the genome hits. This finding may be significant as this repeat sequence may contribute to genetic diversity of the population.

Phylogenetic analysis helped compare the relationship of the proteases encoded on the BAC to one another and to previously identified AC proteases. The phylogram contains additional information about the divergence that has occurred since the proteases shared a common ancestor. From this phylogenetic tree, the previously identified AC-4, AC-5 and newly identified SJ03 and SJ07

have the longest branch lengths indicating that they have undergone more genetic change than some of the others. There is no indication of time on this graph, thus it is not known whether those genes that are more diverse occurred earlier, later or in concurrence with the less diverse genes. It does however suggest that there has been adaptation of the genes within the family in order to maintain a number of different functions. The mechanisms responsible for diversification, specifically the pattern of nucleotide substitution is therefore of interest to examine.

Hughes (1994) described a model depicting the evolution of new proteases, and termed this 'the model of evolution of functionally novel proteins by mutation during non-functionality' or the MDN model. This indicated that after duplication has occurred, one gene copy is redundant, thus any nucleotide changes occurring to this gene will be selectively neutral. This accumulation of amino acid changes may be occurring due to Darwinian selection in which nucleotide changes favour new functions. There is however no definitive evidence of the mechanism of gene duplication. Hughes (1994) outlined a new model with the aim of further understanding the evolution of multigene families. This suggests that during the period before duplication, genes may perform more than one function and thus after duplication each protease would perform one of these functions. Post-duplication, natural selection may favour certain amino acid changes that aid further changes to benefit the protease in its specific function.

In conclusion, bioinformatic annotation of BAC 18f22 has enabled the coding and non-coding regions of the eight tandemly arranged genes to be studied in detail. This showed that the encoded proteases belong to the AC family and identified six additional members. The duplication and divergence of these tandemly arranged genes may be to enable co-ordinated gene expression and/or differences in specificity. At this point, the exact function and importance of *H. contortus* AC cysteine protease genes is unknown; this family is not closely related to the *C. elegans* CPR cathepsin B proteases (35-57% identity) and therefore cannot be studied by reference to *C. elegans*. Currently *sj01* is incomplete and it is unknown if previously identified AC-4 and AC-5 are linked to the other members. Further sequence data as well as improved assembly, should enable the complete characterisation of this cysteine protease gene family in *H. contortus*.

## **Chapter 4**

Other multigene protease families of *H. contortus*  
and analysis of *cpr-6*, a unique protease gene

## 4.1 Introduction

Cathepsin B-like (CBL) cysteine proteases have been identified as potentially important vaccine candidates and for this reason the diversity and possible roles of CBL proteins is of great interest (Jasmer *et al.*, 2004). The work presented in Chapter 3 identified additional members of the previously identified AC protease family (Pratt *et al.*, 1990) and indicated that they arose from gene duplication. Genes encoding related CBL proteases have been previously identified. In *H. contortus* these include the HmCP family (Skuce *et al.*, 1999) containing six members (HmCP1-6), and an additional gene named *gcp-7* (Rehman and Jasmer, 1998). Interestingly, the proteases discussed so far (AC family, HmCP family and GCP-7) do not show strong similarity to *C. elegans* CBL proteases and other related parasitic nematodes, suggesting their functions may be specific to *H. contortus*. An exception to this is the Ce-CPR-6 protease which shows significant identity to a CBL enzyme from *Ascaris suum*, named As-CP-1 (Rehman and Jasmer, 1999). The strong sequence conservation suggests that *cpr-6*-like genes may have a house keeping function and for this reason are well conserved between species. To date, CPR-6-like sequences from other parasitic nematodes have not been well characterised, therefore it is not known how widely conserved this protease is nor whether it is also present as a multigene family.

Thus, a number of *H. contortus* protease genes have been identified, however their genomic organisation and whether they represent different genes or alleles of the same gene is unknown.

The main aims of this chapter were to;

- Identify and annotate *H. contortus* cathepsin B protease genes related to the HmCP gene family as well as to *gcp-7*.
- Annotate and characterise *H. contortus cpr-6* and try to determine the potential conserved role the CPR-6 protease may play in nematodes.

## 4.2 Results

### 4.2.1 Multigene families in *H. contortus*

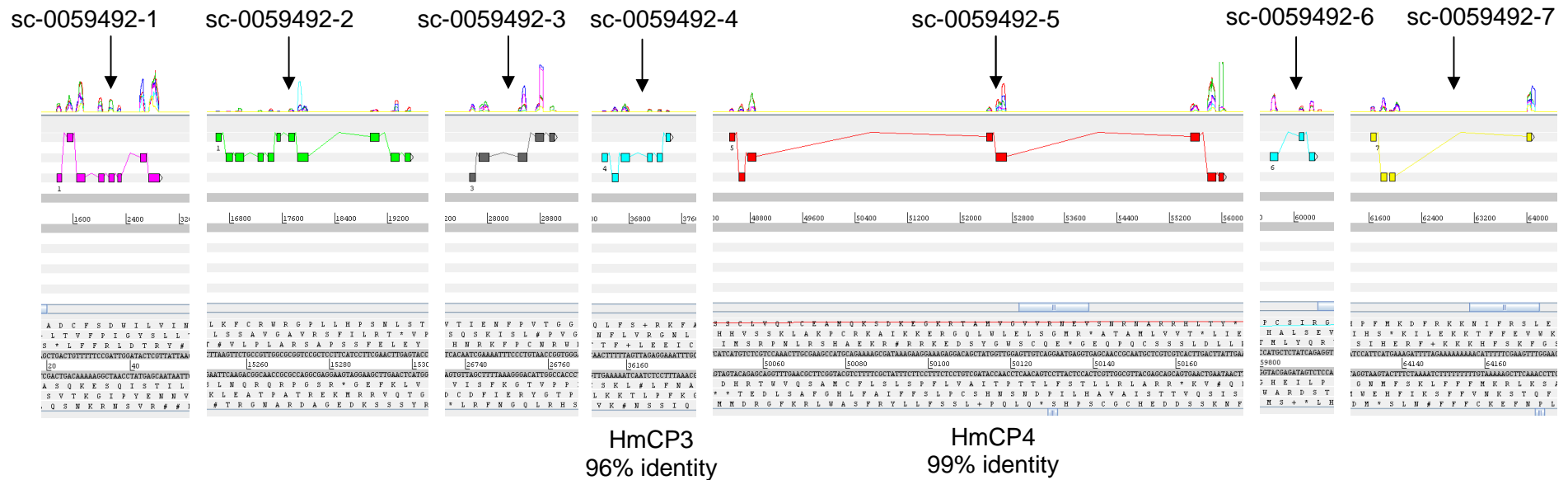
#### 4.2.1.1 *H. contortus* cysteine protease genes related to the HmCP gene family

After the discovery of the tandemly arranged multigene family present on BAC 18f22, work was carried out to determine if other *H. contortus* proteases are members of large gene families and if they are also tandemly arranged. The HmCP gene family was characterised by Skuce *et al.*, (1999) and encodes six cathepsin B cysteine proteases (HmCP1-6). The amino acid sequence of HmCP4 was used to carry out a tBLASTn search against the *H. contortus* assembled BAC, contig and supercontig data (21/08/08)(all reads) on the Sanger webpage. The highest level of similarity was found with supercontig\_0059492 (1.7e-117) and the partial sequence of seven genes on this supercontig was subsequently annotated in Artemis. The genes identified were all very similar to the previously identified *H. contortus* six member HmCP cysteine protease family. Figure 4.1 is an Artemis screenshot indicating the structure of the genes present on the supercontig. *sc-0059492-2* has the most sequence available and from this annotation 11 exons can be identified, it is predicted that the start codon is located on a separate exon as observed with the AC family of protease genes on BAC 18f22. By identifying the conserved GT splice acceptor site sequence present after an ATG sequence, as found for the BAC 18f22 proteases, it was possible to predict the start methionine codon for the five protease genes on supercontig\_0059492 for which the location of the second exon was known. The ATG start codons are located between 72 and 80 bp upstream from the start of the second exon. The exact size and positioning of the genes cannot be confirmed as there are un-sequenced regions within this supercontig and the possibility of assembly errors. One example of this is the fifth gene present on the supercontig. From Figure 4.1 there appears to be two very long introns. However, comparison of the genomic sequence with available cDNA sequence shows that there are three exons missing due to gaps in the sequence within these introns (indicated by Ns in the supercontig sequence).

The percentage identity of the previously identified HmCP cysteine proteases and the proteases encoded by the supercontig was calculated and the results shown in Table 4.1. The sequence of HmCP3 and sc-0059492-4 encoded by the supercontig share 96% amino acid identity and 97% DNA identity, suggesting that they are the same gene. Similarly, HmCP4 and sc-0059492-5 encoded by the supercontig show 99% amino acid identity and 98% DNA identity. This indicates that they are also the same gene, with the sequence changes likely due to variation in the parasite population. From the genomic information available, only four exons for sc-0059492-7 present on the supercontig could be identified. Similarly, for sc-0059492-3 and sc-0059492-6, only five and three exons, respectively could be identified. The sc-0059492-6 and sc-0059492-7 proteases show 79% amino acid identity and 92% DNA identity, and combined with EST data (see later) suggest that they are the same gene. Apart from HmCP3 and HmCP4 none of the other HmCP proteases could be identified from the supercontig data on the Sanger website. The sequence gaps as well as the difficulty in defining closely related genes highlights some of the challenges in trying to annotate the available genome sequence.

# supercontig\_0059492

92 kb 5' → 3'



**Figure 4.1 Annotation of the HmCP-related protease genes on supercontig\_0059492**

Boxes indicate exons and lines introns. The order displayed is the order present on the supercontig, however the precise locations of the genes in relation to each other may not be correct due to unsequenced regions and possible assembly errors. Intron sizes are not exact due to the presence of un-sequenced regions. The full gene sequence is predicted to contain 11 exons (excluding the start codon exon) with only sc-0059492-2 being identified in full, with exons missing from other genes. RNA sequence data is shown at the top and helped confirm correct gene annotation.



| Protease     | HmCP1 | HmCP2 | HmCP3 | HmCP4 | HmCP5 | HmCP6 | sc-0059492-1 | sc-0059492-2 | sc-0059492-3 | sc-0059492-4 | sc-0059492-5 | sc-0059492-6 | sc-0059492-7 |
|--------------|-------|-------|-------|-------|-------|-------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| HmCP1        | 100   | 66    | 44    | 44    | 44    | 39    | 41           | 40           | 36           | 39           | 38           | 41           | 19           |
| HmCP2        |       | 100   | 41    | 40    | 42    | 36    | 38           | 38           | 40           | 39           | 36           | 44           | 13           |
| HmCP3        |       |       | 100   | 68    | 62    | 58    | 65           | 69           | 72           | 96           | 65           | 64           | 65           |
| HmCP4        |       |       |       | 100   | 65    | 60    | 67           | 72           | 69           | 69           | 99           | 63           | 61           |
| HmCP5        |       |       |       |       | 100   | 51    | 61           | 61           | 51           | 66           | 57           | 64           | 37           |
| HmCP6        |       |       |       |       |       | 100   | 75           | 62           | 47           | 66           | 56           | 72           | 58           |
| sc-0059492-1 |       |       |       |       |       |       | 100          | 73           | 27           | 68           | 45           | 68           | 48           |
| sc-0059492-2 |       |       |       |       |       |       |              | 100          | 68           | 70           | 72           | 65           | 61           |
| sc-0059492-3 |       |       |       |       |       |       |              |              | 100          | 0            | 69           | 0            | 73           |
| sc-0059492-4 |       |       |       |       |       |       |              |              |              | 100          | 37           | 64           | 50           |
| sc-0059492-5 |       |       |       |       |       |       |              |              |              |              | 100          | 23           | 60           |
| sc-0059492-6 |       |       |       |       |       |       |              |              |              |              |              | 100          | 79           |
| sc-0059492-7 |       |       |       |       |       |       |              |              |              |              |              |              | 100          |

**Table 4.1 Percentage identity at the amino acid level of proteases encoded by supercontig\_0059492 and the *H. contortus* HmCP family**

Indicated in red are the percentage identities of particular interest. High percentage identities for HmCP3 and sc-0059492-4 and HmCP4 and sc-0059492-5 are shown. In addition to this are the 0% noted for a number of the proteases that do not have overlapping sequence and the high percentages indicated for the short regions of sc-0059492-3 and sc-0059492-6 that overlap with sc-0059492-7. HmCP1 and HmCP2 are more similar to each other than to any of the other proteases encoded by this supercontig.

The expression profiles of the genes on supercontig\_0059492 were analysed. Transcriptome information for the genes has been included in Figure 4.1 using data obtained from adult and infective L3 stages of *H. contortus* worms (R. Laing, J. Gilleard and M. Berry unpublished data). This information helped confirm the correct annotation of the supercontig and also identified that no other genes are present on this supercontig. The transcriptome data indicates whether gene sequences are expressed, however it cannot be used to determine the level of expression. Further to this, Nembase (<http://www.nematodes.org/nembase4/>) was used to search the *H. contortus* EST data to examine transcript abundance. The genomic DNA sequence for all of the genes present on supercontig\_0059492 was used in BLAST search against Nembase, using a 95% sequence identity as a cut off. This analysis identified sc-0059492-5 as having the highest (38) number of associated ESTs, followed by sc-0059492-6 and sc-0059492-7 (28), which are associated with the same Nembase Cluster, then sc-0059492-4, sc-0059492-1 and sc-0059492-3 (Table 4.2). No EST data was identified for sc-0059492-2 which was interesting considering transcriptome data is evident. sc-0059492-6 and sc-0059492-7 show high identity with Nembase cluster HCC00328\_1. Alignment of sc-0059492-6 and sc-0059492-7 sequences with the cluster sequence shows that these are different, but overlapping, regions of the same gene.

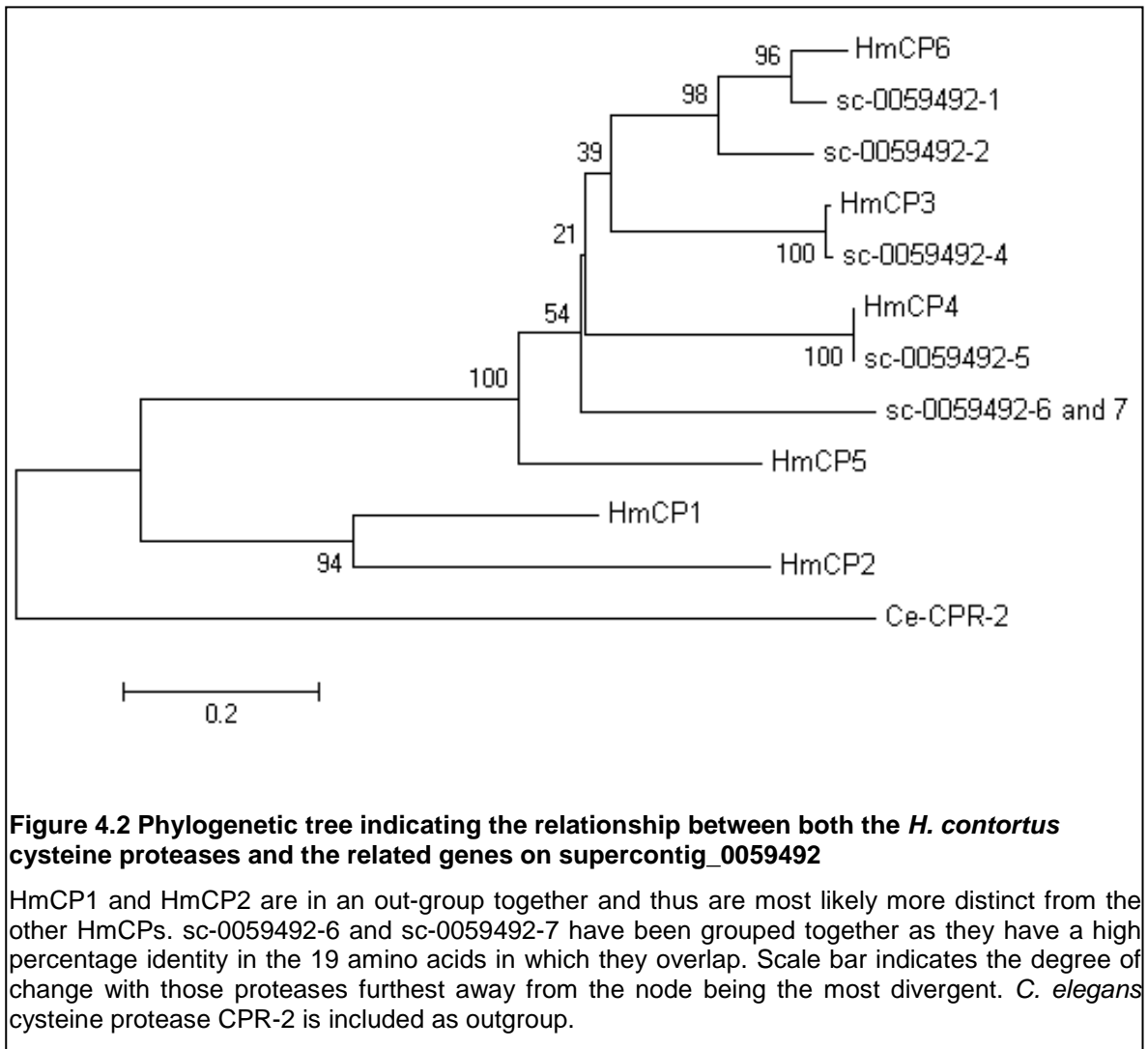
| Gene         | Nembase cluster number | Percentage DNA identity | Sequence coverage (bp) | Number of ESTs |
|--------------|------------------------|-------------------------|------------------------|----------------|
| sc-0059492-1 | HCC06366_1             | 98                      | 625/702                | 15             |
| sc-0059492-2 | -                      | -                       | -                      | -              |
| sc-0059492-3 | HCC03301_1             | 98                      | 516/549                | 3              |
| sc-0059492-4 | HCC01418_2             | 99                      | 394/477                | 23             |
| sc-0059492-5 | HCC00020_2             | 99                      | 546/816                | 38             |
| sc-0059492-6 | HCC00328_1             | 100                     | 254/546                | 28             |
| sc-0059492-7 | HCC00328_1             | 96                      | 144/291                | 28             |

**Table 4.2 EST data for the genes present on supercontig\_0059492**

Information obtained from Nembase BLAST searches using the maximum sequence information available for each gene. All cluster sequences from adult *H. contortus* stage (<http://www.nematodes.org/nembase4/>).

Phylogenetic analysis was used to identify the relationship of the proteases encoded by supercontig\_0059492 to one another and to the HmCP family (Figure

4.2). As expected from the percentage identity, HmCP4 and sc-0059492-5 group together, as do HmCP3 and sc-0059492-4. sc-0059492-3 was not included in phylogenetic analysis as only a short region of sequence is available on the supercontig, and thus no reliable comparison could be carried out. Consistent with the low percentage identity that HmCP1 and HmCP2 share with the HmCP-related proteases, these also group separately by phylogenetic analysis. This raises the question of whether HmCP1 and HmCP2 are members of this family.

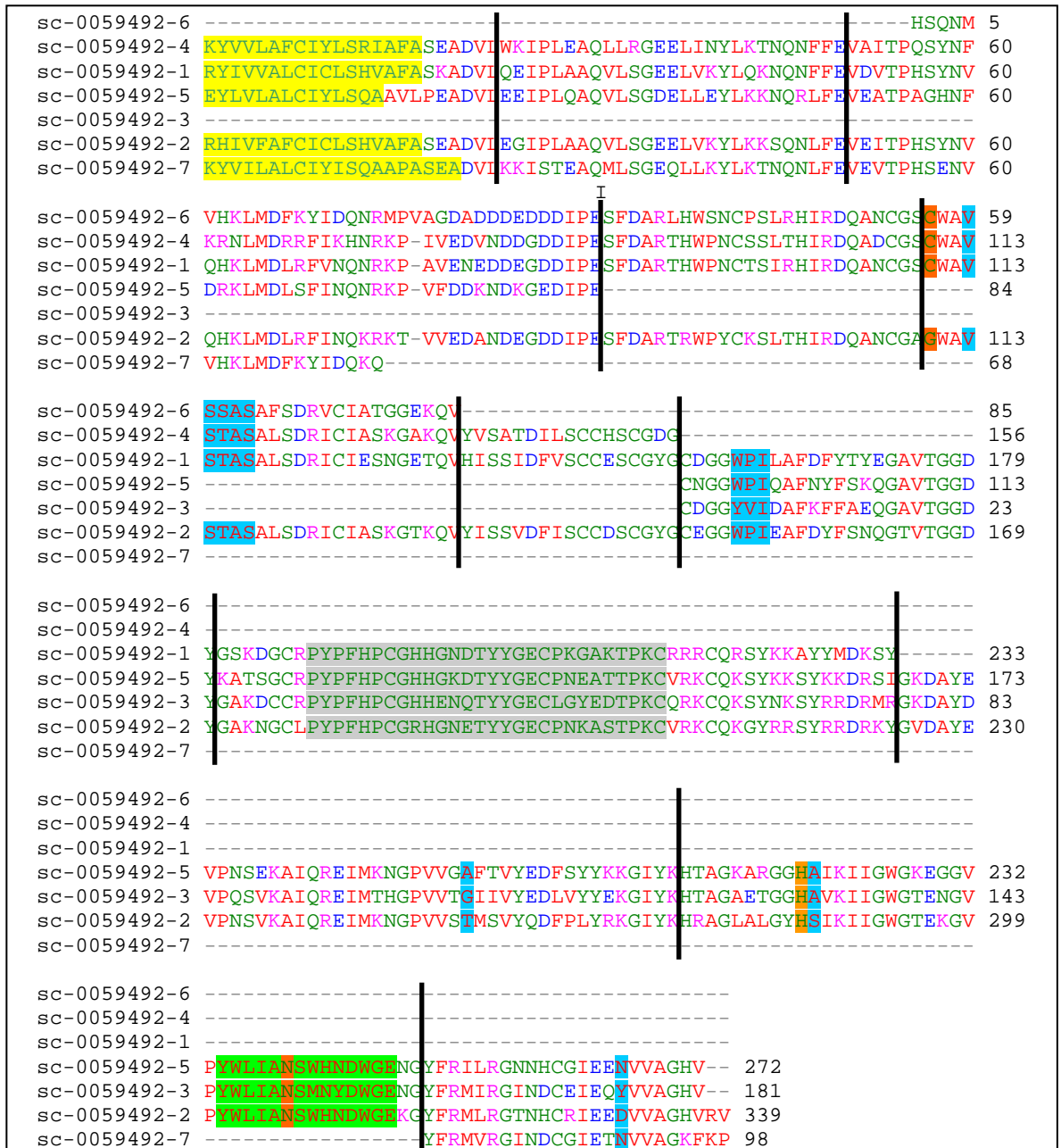


The proteases present on supercontig\_0059492 were further analysed to determine individual relationships. The seven proteases were aligned using Clustal W2 and the results indicated in Figure 4.3. As previously observed for the cysteine proteases encoded by BAC 18f22, intron/exon boundaries are very highly conserved and these are indicated on the alignment.

Where sequence was available, analysis of the active site regions (cysteine, histidine and asparagine) and associated amino acids was carried out. The proteases encoded by this supercontig are all Type B (VSTAA; Chapter 3.1.2.2) as found for the AC proteases. Analysis of HmCP1 and HmCP2 identified that although these are Type B, the middle T is replaced with an A, consistent with them being divergent from the other HmCP proteases. Interestingly, sc-0059492-2 present on the supercontig has a G in place of the cysteine active site and thus is likely to be a pseudogene.

As mentioned in Chapter 3.1.2.2, using a rat cathepsin B sequence as a reference, Rehman and Jasmer (1999) identified S2 subsites Y75, P76, S77, A173, A200 and E245 influencing substrate specificity. The S2 subsites have been identified for each of the HmCP proteases and the results displayed in Table 4.3, with the AC-2 sequence included as a reference. In most instances there is a high level of conservation around the active site regions. One sequence difference to note is that in sc-0059492-2 present on the supercontig there is an S instead of the A at position 287. There is also significant divergence in the conserved WPI S2 subsite sequence at positions 161-163, for HmCP1, HmCP2 and HmCP3, while other HmCPs show conservation with the AC proteases in these positions, suggesting that they may have similar specificities.

In addition to the conserved active and S2 subsites, there is a proposed haemoglobinase motif (YWLVANSW--DWGD) located around the region of the asparagine active site (306) (Baig *et al.*, 2006). This motif is present in all the HmCP proteases identified; however there are a number of sequence variations within this family, particularly in HmCP1, 2 and 5, again suggesting these may have specificities and functions different to the others.



**Figure 4.3 Amino acid alignment of the proteases on supercontig\_0059492**

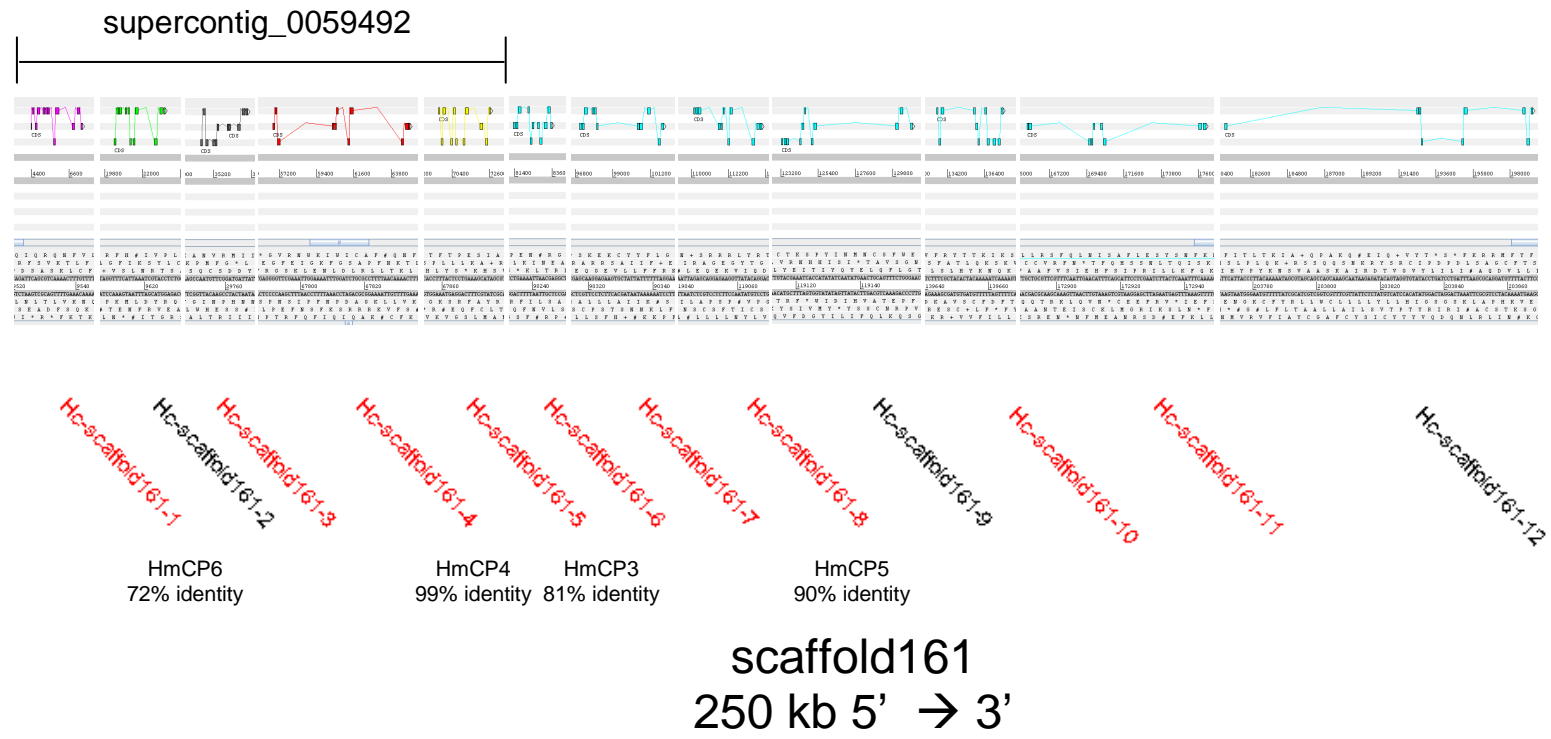
The signal peptide region is highlighted in yellow, the three active site regions in orange and the S2 subsites in blue. The occluding loop region is highlighted in grey and the haemoglobinase motif indicated in green. Intron/exon boundaries are indicated by the black lines and un-sequenced regions by dashed lines. The pro-peptide cleavage site is indicated by I.

| Sequence     | Cysteiny active site signature sequence (113-117) | S2 Subsite        |         |         |         | Haemoglobinase motif (301-315) |
|--------------|---|-------------------|---------|---------|---------|--------------------------------|
|              |   | Y, P, S (161-163) | A (251) | A (287) | E (331) |                                |
| AC-2         | VSTAA   | WPI               | S       | A       | T       | FWLIANSW--DWGE                 |
| HmCP1        | VSAAE   | MDH               | A       | A       | -       | YWNVANSW--DW--                 |
| HmCP2        | VSAAS   | YDH               | A       | A       | V       | YWTVANSW--DWGG                 |
| HmCP3        | VSTAS   | YVI               | A       | A       | N       | YWIIANSW--DWGE                 |
| HmCP4        | VSTAS   | WPI               | A       | A       | N       | YWLIANSW--DWGE                 |
| HmCP5        | VSTAA   | WPI               | A       | A       | -       | YWIVKNSW--DW--                 |
| HmCP6        | VSTAS   | WPI               | A       | A       | E       | YWLIANSW--DWGE                 |
| sc-0059492-1 | VSTAS   | WPI               | -       | -       | -       | -                              |
| sc-0059492-2 | VSTAS   | WPI               | T       | S       | D       | YWLIANSW--DWGE                 |
| sc-0059492-3 | -   | YVI               | G       | A       | Y       | YWLIANSM--DWGE                 |
| sc-0059492-4 | VSTAS   | -                 | -       | -       | -       | -                              |
| sc-0059492-5 | -   | WPI               | A       | A       | N       | YWLIANSW--DWGE                 |
| sc-0059492-6 | VSSAS   | -                 | -       | -       | -       | -                              |
| sc-0059492-7 | -   | -                 | -       | -       | N       | -                              |

**Table 4.3 Conserved regions within the *H. contortus* cysteine proteases and the proteases present on supercontig\_0059492**

AC-2 has been added to the table as a reference to the previously characterised AC protease family and those proteases identified on BAC 18f22. Missing sequence is indicated by dashed lines.

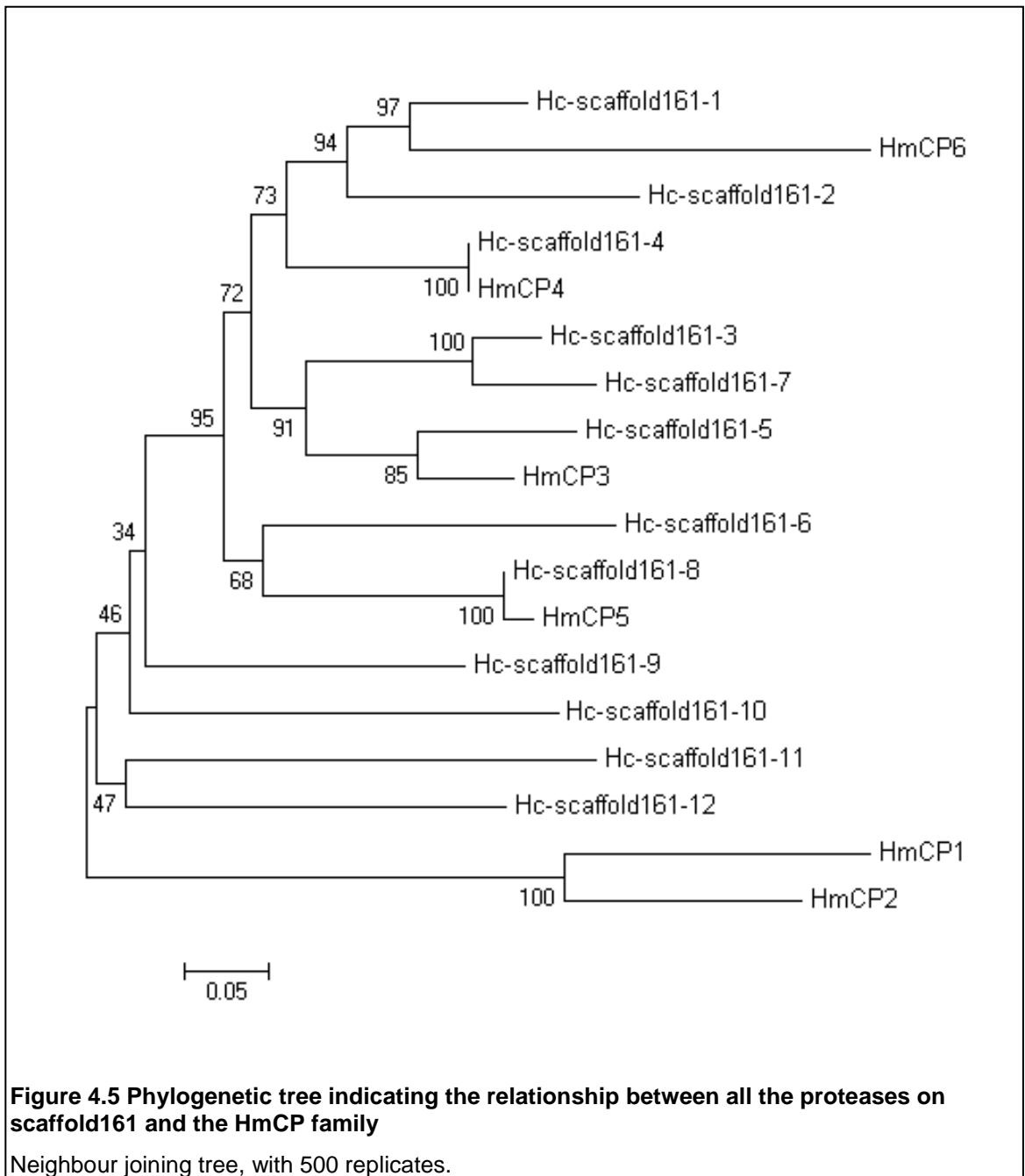
New *H. contortus* sequence data has recently been generated and assembled at the Wellcome Trust Sanger Institute and is available on the Sanger ftp site <ftp://ftp.sanger.ac.uk/pub/pathogens/Haemonchus/contortus/genome/> (folder12062012, James Cotton unpublished data). A BLASTn search against the haem\_supercontigs.12062012 file using the supercontig\_0059492 sequence identified high similarity with scaffold161. This scaffold is almost 250 kb in length and thus provided extra information about the sequence surrounding this supercontig. sc-0059492-3 and sc-0059492-4 for which the end and start, respectively, were known have been re-positioned and combined to give a complete protease sequence. There is additional sequence added between sc-0059492-6 and sc-0059492-7, providing complete protease sequence and confirming these as one gene. Sequence information that was previously missing for a number of the proteases is now present and it is likely that there are five complete proteases from the partial proteases initially identified on the supercontig. Also encoded on this scaffold are an additional seven proteases that are related to the HmCPs. One of these new proteases has a high percentage identity, 90% AA with HmCP5. The newly completed sc-0059492-1 present on the supercontig has a 72% AA identity with HmCP6. This suggests that all the highly similar HmCP proteases previously published are closely located in the genome. Figure 4.4 is an Artemis screenshot of scaffold161 indicating the structure of the genes present, how the genes previously present on supercontig\_0059492 have been reorganised and the names for these genes. Figure 4.5 is a phylogenetic tree indicating the relationship between all the proteases on scaffold161 and the HmCP family. The tree indicates that on the scaffold there are proteases with a high percentage identity to all of the HmCPs except HmCP1 and HmCP2. It also indicates that the Hc-scaffold161-3 sequence is not the most similar to HmCP3, this would have been expected given the high percentage identity of the original partial sc-0059492-4 on the supercontig. Instead with more sequence data there is a higher similarity of Hc-scaffold161-5 to HmCP3.



**Figure 4.4 Annotation of the CBL protease genes on scaffold161 (12062012 file)**

The order displayed is the order of proteases on the scaffold. The scaffold sequence obtained from the ftp site has been reversed so that proteases are displayed on the positive strand in the forward orientation. The region covered by supercontig\_0059492 is indicated, with 5 complete proteases in this region. Of the 12 proteases identified on the scaffold, those highlighted in red have a high identity with previously published partial and complete gene sequences. The accession numbers for these sequences are in Appendix 3 Table 1.





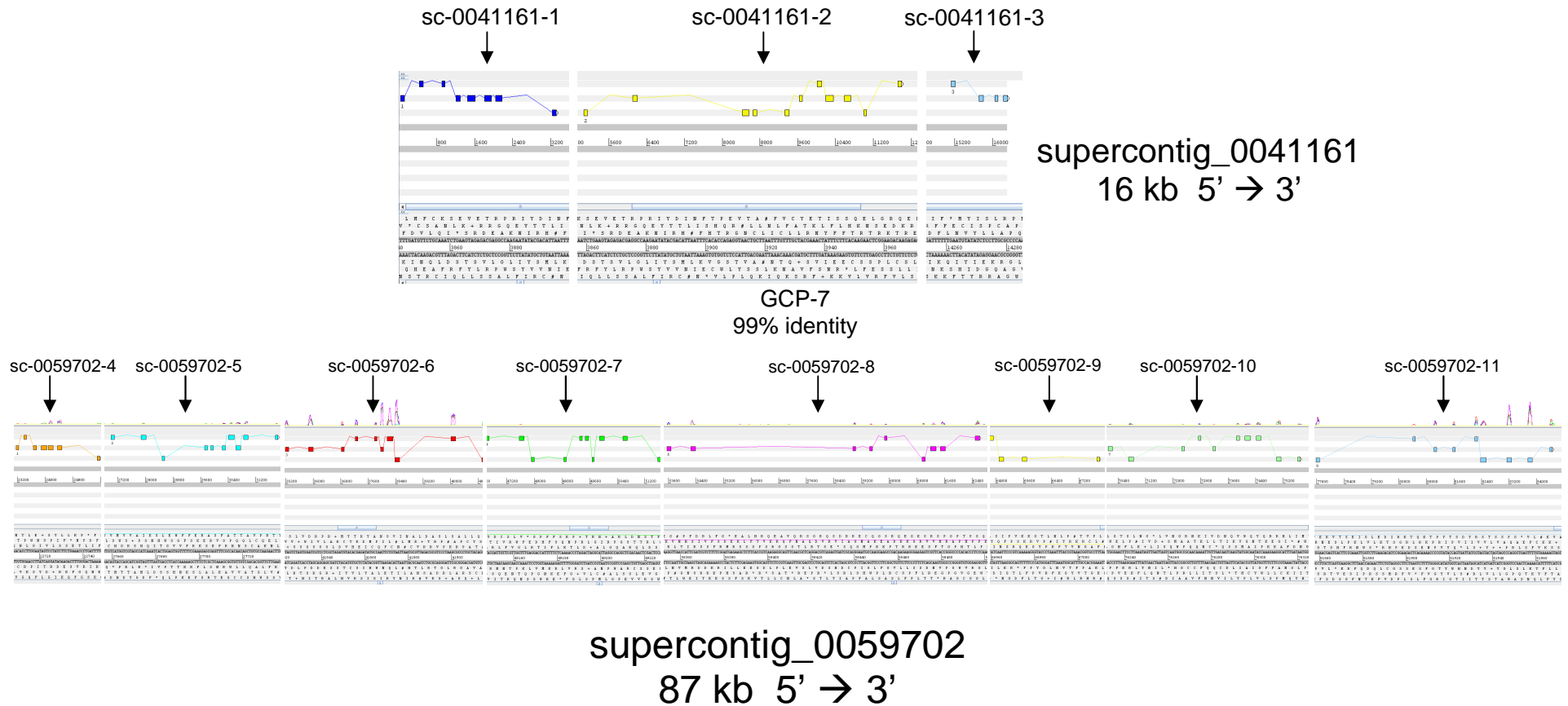
The sequence information gathered from the annotation of supercontig\_0059492 and scaffold161 shows that there are additional members to the already identified *H. contortus* HmCP cysteine protease family. It also suggests that those HmCP proteases which are most similar are situated closely in the genome, suggesting that like the AC family, these have undergone recent duplication and sequence divergence.

#### 4.2.1.2 A *H. contortus* multigene family related to *Hc-gcp-7*

Rehman and Jasmer (1998) sequenced protease genes from *H. contortus* gut extracts. One sequence identified was similar, but distinct from the already characterised *H. contortus* AC and HmCP families, and was designated *gcp-7* (AF046229). To identify any related sequences a tBLASTn search using *H. contortus gcp-7* sequence, was carried out against the *H. contortus* supercontig data (21/08/08)(all reads). Two supercontigs were identified with high similarity to this sequence; supercontig\_0041161 and supercontig\_0059702.

Initial analysis of supercontig\_0041161 (16 kb) identified three genomic regions that had high sequence similarity to *gcp-7*. Of the three genes identified on this supercontig, the first was missing the start of the sequence and the third was missing the end, as they were not included on this supercontig. The third *gcp-7* related gene was additionally missing the first four exons. The second gene present on the supercontig was the only gene for which almost the full sequence was available, missing only the first two exons and part of the second last exon. Analysis of supercontig\_0059702 identified eight genomic regions with high sequence similarity to *gcp-7*. These genes were numbered 4-11, to continue chronologically from the genes identified on supercontig\_0041161. The first two exons were not able to be identified for any of the genes, most likely due to sequence divergence.

Figure 4.6 is an Artemis screenshot indicating the structure of the genes present on supercontig\_0041161 and supercontig\_0059702. The exact location and size of the genes cannot be confirmed as there are un-sequenced regions (represented by Ns) within the supercontigs and potential assembly errors. The percentage identity of the proteases encoded by the two supercontigs was calculated and the results indicated in Table 4.4. GCP-7 and sc-0041161-2 share a 99% amino acid identity and 97% DNA identity, suggesting that they represent the same gene. The sequences for proteases sc-0059702-8 and sc-0059702-10 are almost complete, with nine and ten exonic sequence regions available, respectively. These proteases share an 89% identity at the amino acid level and a 93% identity at the DNA level. Based on EST evidence (see later EST data) it is likely that sc-0059702-8 and sc-0059702-10 represent the same gene and have been positioned separately on the supercontig due to assembly errors.



**Figure 4.6** Annotation of the protease genes on supercontig\_0041161 and supercontig\_0059702

Boxes indicate exons and lines introns. The order displayed is the order present on the supercontig, however the locations of the genes in relation to each other may not be correct due to assembly errors. Intron sizes are not exact due to the presence of un-sequenced regions. The full gene sequence contains 11 exons (excluding the start codon exon) and it can be observed for the genes present on these supercontigs that there are a number of exons missing from a number of genes. RNA sequence data is shown at the top and helped confirm correct gene annotation.

| Protease      | sc-0041161-1 | sc-0041161-2 | sc-0041161-3 | sc-0059702-4 | sc-0059702-5 | sc-0059702-6 | sc-0059702-7 | sc-0059702-8 | sc-0059702-9 | sc-0059702-10 | sc-0059702-11 | GCP-7 |
|---------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|---------------|---------------|-------|
| sc-0041161-1  | 100          | 51           | 68           | 64           | 64           | 65           | 55           | 60           | 58           | 65            | 63            | 59    |
| sc-0041161-2  |              | 100          | 69           | 51           | 56           | 57           | 49           | 49           | 46           | 57            | 52            | 99    |
| sc-0041161-3  |              |              | 100          | 48           | 72           | 70           | 55           | 68           | 0            | 67            | 64            | 55    |
| sc-0059702-4  |              |              |              | 100          | 77           | 74           | 59           | 64           | 69           | 69            | 69            | 58    |
| sc-0059702-5  |              |              |              |              | 100          | 68           | 68           | 64           | 66           | 65            | 69            | 57    |
| sc-0059702-6  |              |              |              |              |              | 100          | 65           | 65           | 64           | 67            | 69            | 59    |
| sc-0059702-7  |              |              |              |              |              |              | 100          | 72           | 73           | 77            | 67            | 55    |
| sc-0059702-8  |              |              |              |              |              |              |              | 100          | 66           | 89            | 73            | 54    |
| sc-0059702-9  |              |              |              |              |              |              |              |              | 100          | 73            | 70            | 56    |
| sc-0059702-10 |              |              |              |              |              |              |              |              |              | 100           | 78            | 57    |
| sc-0059702-11 |              |              |              |              |              |              |              |              |              |               | 100           | 59    |
| GCP-7         |              |              |              |              |              |              |              |              |              |               |               | 100   |

**Table 4.4 Amino acid identities of GCP-7 and the proteases encoded by supercontig\_0041161 and supercontig\_0059702**

Indicated in red are the high percentage identities for GCP-7 & sc-0041161-2, and sc-0059702-8 and sc-0059702-10. In addition to this is the 0% noted for the third and ninth proteases with only partial sequences available.

Transcriptome data is available for supercontig\_0059702 and has been included in Figure 4.6. From this information it is evident that for a number of the genes there is a very low signal, with only *sc-0059702-4*, *sc-0059702-6*, *sc-0059702-8*, *sc-0059702-10* and *sc-0059702-11* having easily identifiable peaks. The Nembase webpage was also used to search the *H. contortus* adult stage EST data for expression of all of the *gcp-7* related sequences, using BLAST analysis. A 95% sequence similarity was used as a cut off when identifying potential matches. No EST data was obtained for *sc-0059702-5*, *sc-0059702-7* and *sc-0059702-9*, indicating that they may either not be expressed or be expressed at a low level. *sc-0059702-8* and *sc-0059702-10* have the highest number of associated ESTs and are both associated with the same Nembase Cluster indicating, as mentioned previously, that they may be the same gene (Table 4.5).

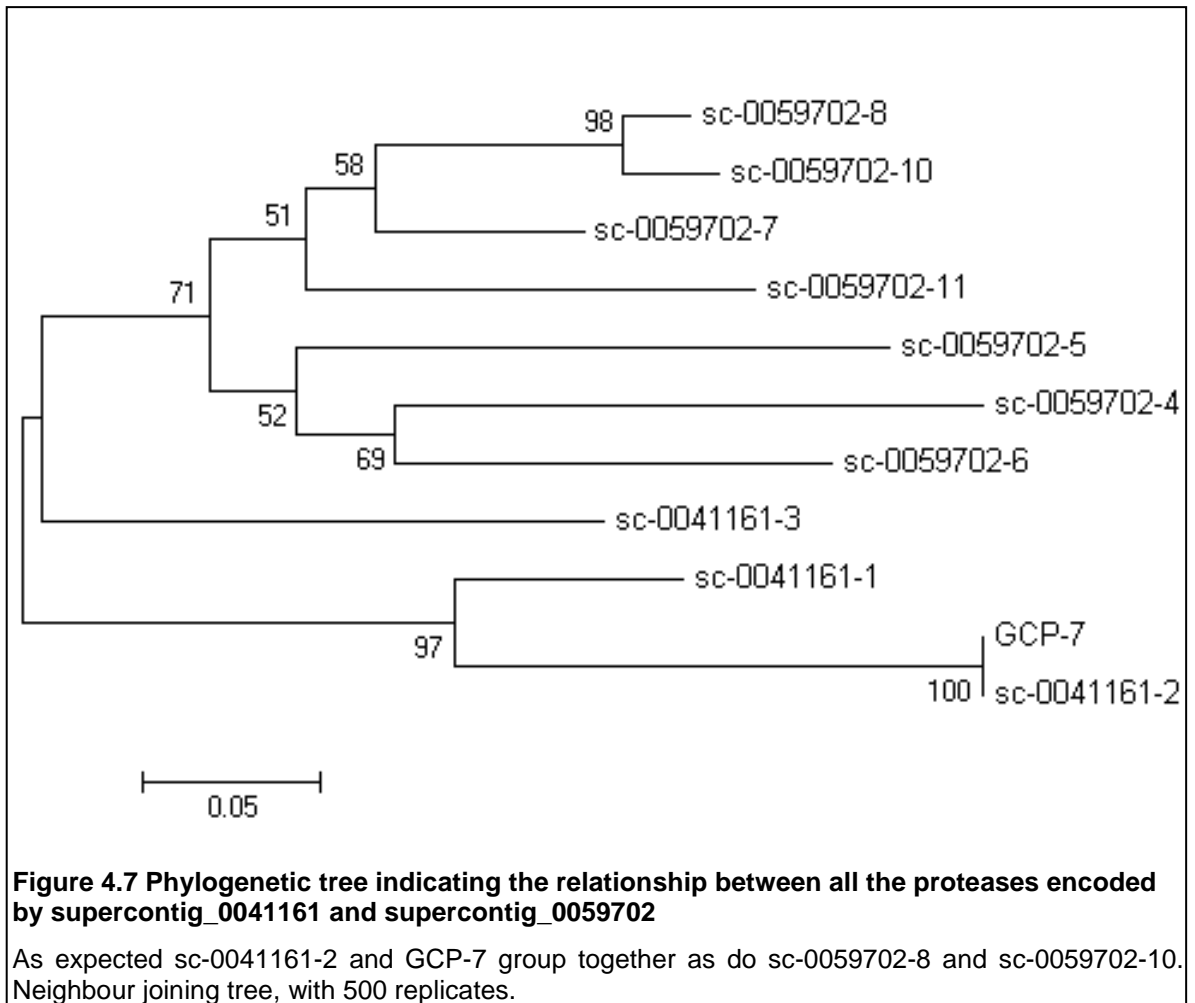
In addition to providing information about expression, the EST data was used to identify the second exon of a number of the proteases however the start methionine codon was unable to be identified due to it being on a separate exon. The Nembase protease sequences were used to carry out tBLASTn analysis against the *H. contortus* supercontig data (21/08/08)(all reads) on the Sanger webpage and the second exon was identified for all proteases except *sc-0041161-1*, *sc-0059702-4* and *sc-0059702-9*. The complete sequence for *sc-0041161-3* on the supercontig was identified from the Nembase EST data. Thus on analysis of the supercontig, the second exon could be identified, however downstream of this there is an un-sequenced region between this exon and the latter exons. The full Nembase sequence was aligned with the other proteases to ensure that there was no duplication or assembly error; this protease was confirmed to be a separate gene. The Nembase sequence that was absent on the supercontig was used in tBLASTn against the *H. contortus* supercontig data (21/08/08)(all reads) on the Sanger webpage in an attempt to identify the rest of the protease at a different location within the *H. contortus* genome, however no matches were identified.

| Gene                 | Nembase Cluster number | Percentage DNA identity | Sequence coverage (bp) | Number of ESTs |
|----------------------|------------------------|-------------------------|------------------------|----------------|
| <i>sc-0041161-1</i>  | HCC00986_1             | 96                      | 579/756                | 1              |
| <i>sc-0041161-2</i>  | HCC00166_1             | 97                      | 877/969                | 6              |
| <i>sc-0041161-3</i>  | HCC03318_1             | 95                      | 263/273                | 3              |
| <i>sc-0059702-4</i>  | HCC00310_2             | 97                      | 391/684                | 9              |
| <i>sc-0059702-5</i>  | -                      | -                       | -                      | -              |
| <i>sc-0059702-6</i>  | HCC01012_1             | 97                      | 960/981                | 6              |
| <i>sc-0059702-7</i>  | -                      | -                       | -                      | -              |
| <i>sc-0059702-8</i>  | HCC01173_3             | 98                      | 409/933                | 10             |
| <i>sc-0059702-9</i>  | -                      | -                       | -                      | -              |
| <i>sc-0059702-10</i> | HCC01173_2             | 96                      | 602/1002               | 10             |
| <i>sc-0059702-11</i> | HCC01124_1             | 95                      | 565/852                | 3              |

**Table 4.5 EST data for the genes present on supercontig\_0041161 and supercontig\_0059702**

Information obtained from Nembase BLAST searches using the maximum sequence information available for each gene. All cluster sequences from adult *H. contortus* stage (<http://www.nematodes.org/nembase4/>).

Phylogenetic analysis was used to identify the relationship between the proteases encoded by the two supercontigs. Figure 4.7 is a phylogram which contains additional information about the divergence of the proteases with a scale bar to indicate the degree of change. Those proteases furthest away from the node are the most divergent. *sc-0059702-9* has been removed from the analysis due to low sequence coverage. As expected from analysis of the percentage identity, GCP-7 and *sc-0041161-2* group together. In addition to this, *sc-0059702-8* and *sc-0059702-10* group together, further supporting the hypothesis that these are the same genes.



The proteases present on the supercontigs were aligned using Clustal W2 and the results shown in Figure 4.8. The highly conserved intron/exon boundaries are indicated as are the active site regions, S2 subsites and haemoglobinase motif (Figure 4.8 and Table 4.6). One notable difference between these proteases and those previously described is around the cysteinyl active site. The GCP-7 like proteases are all Type B, however the sequence changes in the motif are the same as the sequences identified in HmCP1 and HmCP2. For all of the GCP-7 related proteases, T129 has been replaced by an A (VSAAE). Interestingly, GCP-7 differs from the others by having VSAAQ rather than VSAAE, which may influence activity. Within the proposed haemoglobinase motif, the YWTV sequence of the GCP-7-related proteases also occurs in HmCP2, it is unknown as to whether this change would affect activity and/or specificity.

|               |  |     |
|---------------|--|-----|
| sc-0059702-4  | -----  | 0   |
| sc-0059702-5  | --VLLIFTSILFHDSFEKVLTIIEFFADRPPIPKYAEQLSGEALVDYVNRQQPFFFAEYLP  | 38  |
| sc-0059702-7  | --VSLIFVLFVSVNSIQRELTAEFAAQPIPMHAQELTGEALVEYVNEKQSYFFAEYYPE    | 38  |
| sc-0059702-9  | -----  |     |
| sc-0059702-8  | --VSLIFITFFVSNLSQRELTVEEFAAQPIPKYAEELTGKALEEYVNNKQSFFKQAKYSE   | 43  |
| sc-0059702-10 | --VSLIFITLFFVSKSLQRELTVEEFAAQPIPKYAEELTGKALEEYVNNKQSFFKAEYSAE  | 43  |
| sc-0059702-11 | --VPLILVIFVISTSSQRGPSTEEFAAQPIPKYAEELTGKALEEYVNTKQSYFK-----    | 33  |
| sc-0059702-6  | --ALLIFCAFVVRSERTVLTAEQFAAKPISKEAQKLTGKALVDYVNEQQSFFKAEYSPD    | 38  |
| GCP-7         | MLVLLVLLSFFTVSSSQKFTRIEEFLAQPITKEAEQLTGEALVEYVNNKQSFFKAKYSP    | 60  |
| sc-0041161-2  | --VLLVLLSFFTVSSSQKFTRIEEFLAQPITKEAEQLTGEALVEYVNNKQSFFKAKYSP    | 38  |
| sc-0041161-3  | --VLLAFSFLLVIHSSYATLTI-----                                    |     |
| sc-0041161-1  | -----  |     |
| sc-0059702-4  | -----  |     |
| sc-0059702-5  | VAEKRLGSLMKMDFLLLPAGMDNVTMVGEPVTNEELPESFDSREKWKDCP-SISYIRDQS   | 97  |
| sc-0059702-7  | VAEKRLGSLMKMEYLRSPGGEYLAMMLEESNAKEEIPESFDAREKWKNT-SIGYIRDQS    | 97  |
| sc-0059702-9  | -----  | 0   |
| sc-0059702-8  | VAEKRLNNLMKMEFLHAPPGEYLTMMPEELDTNQALPESFDARDKWKNCSSVIGYIRDQS   | 103 |
| sc-0059702-10 | VTEKRLNNLMKMEFLHASPGERLTMMPEELDTNEVIPESFDARDKWKNCSSVIGYIRDQS   | 103 |
| sc-0059702-11 | -----WKNCT-SIDYIRDQS   | 47  |
| sc-0059702-6  | VIEQRKRTLKMKELLEHPMQKEIVAKAKKLVINEDIPEESFDAREKWKDCP-SIRYIRDQS  | 97  |
| GCP-7         | VVKRRRQFLPKPQFIERSYNQENVLPPIANITSNDDIPEESFDSREKWKDCP-SLRVIPDQS | 119 |
| sc-0041161-2  | VVKRRRQFLPKPQFIERSYNQENVLPPIANITSNDDIPEESFDSREKWKDCP-SLRVIPDQS | 97  |
| sc-0041161-3  | -----SFDSRQKWKDCP-SIRDYIRDQS                                   | 21  |
| sc-0041161-1  | -----SFDSRDQWKDCP-SLRYIRDQT                                    | 21  |
| sc-0059702-4  | --GSCWAVSAAETMSDRLCIHSNSTLQTLISDTDLLSCCGSSCGHCCEGGSIYAWVYAK    | 58  |
| sc-0059702-5  | NCGSCWAVAAASTMSDRLCIQSKGFQTLISDTDILSCCQPLCGDCNCGSLMRAWYYAR     | 157 |
| sc-0059702-7  | NCGSCWAVSAAETMSDRLCIHTYKGLQTLISDTDILSCCGTYCYGCEGGYAIRAWGFAR    | 157 |
| sc-0059702-9  | -----  | 0   |
| sc-0059702-8  | NCGSCWAVSAAETMSDRLCIGTNGFLQTLISDTDILSCCGMFCGDCCEGGYTIIRAWGYAR  | 163 |
| sc-0059702-10 | NCGSCWAVSAAETMSDRLCIGTNGFLQTLISDTDILSCCGMFCGDCDGGYAIRAWGYAK    | 163 |
| sc-0059702-11 | NCGSCWAVSAAETMSDRVCIATDGLQRRIISDTDILSCCGIYCGFCCEGGYSIRAWSYAR   | 107 |
| sc-0059702-6  | NCGSCWAVSAAETMSDRLCIHSNGTFQTMISDTDMLSCCGLACGFCDDGGLAIGAWFYAQ   | 157 |
| GCP-7         | NCGSCWAVSAAQCMSDRLCIHSQGFKKVLLSATDILACCGKFCGYCDDGGYNARAWKWAT   | 179 |
| sc-0041161-2  | NCGSCWAVSAAQCMSDRLCIHSQGFKKVLLSATDILACCGKFCGYCDDGGYNARAWKWAT   | 157 |
| sc-0041161-3  | NCGSCWAVSAAETMSDRACIHSKGFVKVL-SDTDILSCCGEFCGYR-----            | 66  |
| sc-0041161-1  | KCGSCWAVSAAETMSDRLCIHTQGFVKVMLSDTDILACCGKFCGYCCEGGYNGRAWKWAT   | 81  |
| sc-0059702-4  | KHGVCSSGRYGAEKVCKPYVFHPCGRHQGQKYYGECPK-HMYKTPVCKRYCQYGYGKRYE   | 117 |
| sc-0059702-5  | DHGVCSSGRYEEKGVCKPYAFHPCGRHKGQKYHGECPR-HIYKTPVCKPYCQYGYGKRYK   | 216 |
| sc-0059702-7  | DSGVCSSGRYETI-----VYCQYGYGKRYK                                 | 182 |
| sc-0059702-9  | -----DNCKPYVFQPCGWHSGHKYYGECPYDHTYATPPCKEYCYGYGKRYK            | 47  |
| sc-0059702-8  | DSGVCSSGRYDITGNCKPYVFHPCGFDHGQKYYGMCPRDHTFKTPVCKKYCYGYGKRYN    | 223 |
| sc-0059702-10 | DSGVCSSGRYDITGNCKPYVFPYPCGFNEGQKAYGVCPRDHAYKTPVCKKYCYGYGKRYN   | 223 |
| sc-0059702-11 | DKGVCSSGRYESTGNCKPYVFHPCGRHAGQKFGYDCPRDHLFKTPVCKNYCYGYGKRYK    | 167 |
| sc-0059702-6  | DFGVCSSGRYEEKGVCKPYVFHPCGFDHGQKYYGQCPK-HVYKTPVCKSYCQYGYGKRYQ   | 216 |
| GCP-7         | IAGVVTGGAYKEKGNCKPYVFPQCGAHKGKAFN-NCPS-HPYATPACKPYCQYGYGKRYE   | 237 |
| sc-0041161-2  | IAGVVTGGAYKEKGNCKPYVFPQCGAHKGKAFN-NCPS-HPYATPACKPYCQYGYGKRYE   | 215 |
| sc-0041161-3  | -----  | 66  |
| sc-0041161-1  | ISGVVSSGRYGEKGVCMYVFHPCGSHKNQRFYGCPT-HSYRTPACKPYCQYGYGKRYM     | 140 |
| sc-0059702-4  | NDKFIYAKAVYGIF-SFEPAIQMNIMKNGPVQAAFVYEDFAYYKSGVYVHTAGKDTGGHA   | 176 |
| sc-0059702-5  | DDKFIYAKAVYGIF-SFEDAIRMIIMKKGVPVSAAFVYEDFAYYRGGVYVHTAGKKTGRHA  | 275 |
| sc-0059702-7  | DDKFFVKGAFILP-QNERVIQSQIMKRGVPVQAAFIVYDDFSYYRGGVYVHTAGKARGAHA  | 241 |
| sc-0059702-9  | DDKFFVEGAFTLP-QNERVIQSQIMKRGVPVQAAFIVYDDFSYYKSGVYVHTAGKARGAHA  | 106 |
| sc-0059702-8  | RDKFFAKGAYMLP-QNEALIQSQIMKRGVPVQAAFVYEDFSSYKSGIYVHTAGKDRGAHA   | 282 |
| sc-0059702-10 | RDKVFAKAYMLP-QNEALIQSQIMKRGVPVQAAFVYEDFGAYKSGIYVHTAGKDRGAHA    | 282 |
| sc-0059702-11 | LDKVFPAHKAYILP-EHEGAIKEQIMTKGPVQAAFVYEDFSLYKGGVYVHTAGKSRGAHA   | 226 |
| sc-0059702-6  | ADRVFAKVVYGLY-KDEDIIRMDIMKKGVPVQAAFVYEDFDFYSYKGGVYVHTAGEQNLHA  | 275 |
| GCP-7         | NDKIKARTWYWLP-NDERTIQLEIMQKGPVHAFTNIYEDFEHYEGGVYIHTAGAMEGGHS   | 296 |
| sc-0041161-2  | NDKIKARTWYWLP-NDERTIQLEIMQKGPVHAFTNIYEDFEHYEGGVYIHTAGAMEGGHS   | 274 |
| sc-0041161-3  | -----  |     |
| sc-0041161-1  | KDKVFAKTWYWLPQKDEEAIKAEIFQKGPVHAFTNVYEDFASYKGGVYIHTAGKMKGGHS   | 200 |



|               |              |       |             |       |                    |         |       |
|---------------|--------------|-------|-------------|-------|--------------------|---------|-------|
| sc-0059702-4  | VKIIGWGVENG  | TK    | YNTVANSWNTD | WGE   | NGGYFRILRGENHCGIES | QVFAGDF | 228   |
| sc-0059702-5  | VKIIGWGVQNG  | TK    | YNTVANSWNTD | WGED  | DGGYFRILRGKKHCGIES | SIYTGDF | 327   |
| sc-0059702-7  | VKVIIGWGVQNG | TK    | YNTVANSWNTY | WGED  | EGGYFRILRGNHCFES   | MIAGDF  | 293   |
| sc-0059702-9  | VKVIIGWGVQNG | TK    | YNTVANSWNTY | WGED  | EDGYFRILRGMNHCFES  | MIAG--  | 156   |
| sc-0059702-8  | VKVIIGWGVENG | TK    | YNTVANSWNTD | WGEK  | -----              | -----   | 311   |
| sc-0059702-10 | VKVIIGWGVENG | TK    | YNTVANSWNTD | WGEN  | GGYFRILRGNHCEIES   | IMVAGTF | 334   |
| sc-0059702-11 | VKVIIGWGVENG | TK    | YNTVANSWNTD | WGEN  | GGYFRILRGNHCGIEG   | IMVAGEF | 278   |
| sc-0059702-6  | VKIIGWGVENG  | TK    | YNTVANSWNTD | WGED  | DGGYFRFLRGNHCSIEG  | VLAGDF  | 327   |
| GCP-7         | IKIIGWGVDK   | GVK   | YNTVANSWST  | WGED  | DGGYFRVVRGINNCDIEG | VLAGTF  | 348   |
| sc-0041161-2  | IKIIG        | ----- | -----       | ----- | GYFRVVRGINNCDIEG   | VLAGTF  | 302   |
| sc-0041161-3  | -----        | ----- | -----       | ----- | -----              | -----   | ----- |
| sc-0041161-1  | VKIIGWGVENG  | TK    | YNTVANSWST  | WGEN  | GGYFRVVRGIDNCYIES  | VLAGTF  | 252   |

**Figure 4.8 Amino acid alignment of all the proteases in the GCP-7-like family**

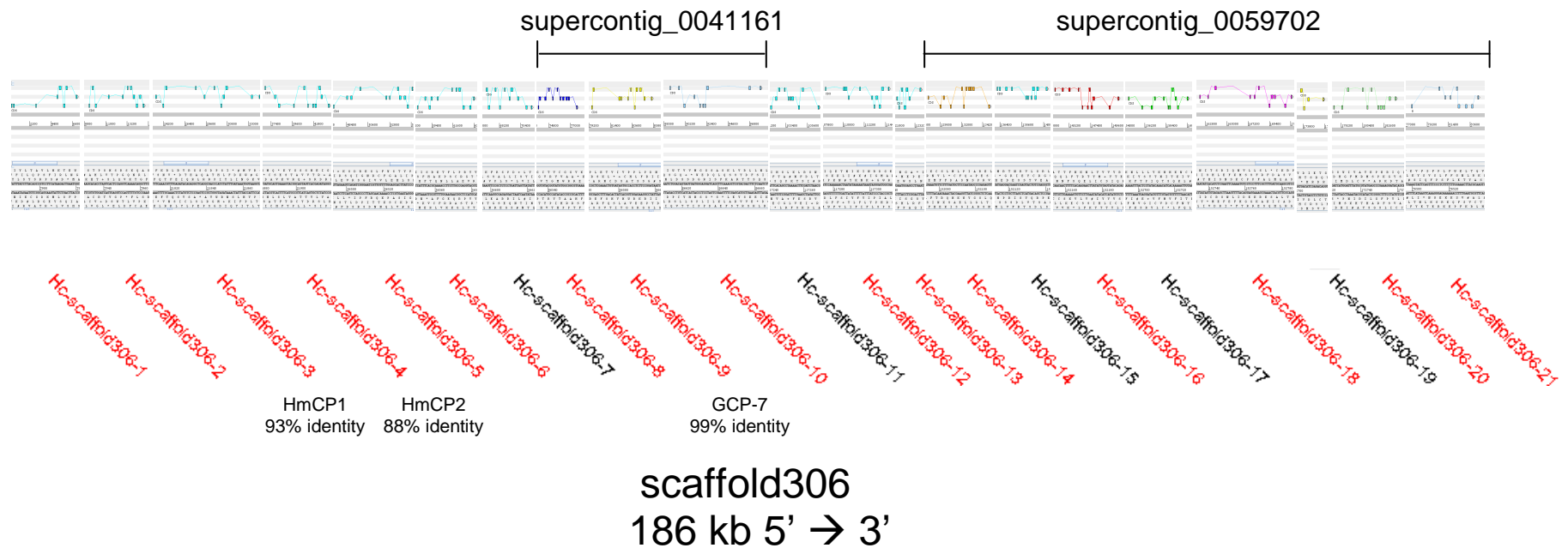
The signal peptide region is highlighted in yellow the three active site regions in orange and the S2 subsites in blue. The occluding loop region is highlighted in grey and the haemoglobinase motif indicated in green. Intron/exon boundaries are indicated by the black lines and un-sequenced regions by dashed lines. The pro-peptide cleavage site is indicated by |.

| Sequence      | Cysteiny active site signature sequence (127-131) | S2 Subsite        |         |         |         | Haemoglobinase motif (310-323) |
|---------------|---|-------------------|---------|---------|---------|--------------------------------|
|               |   | Y, P, S (170-172) | A (269) | A (296) | E (342) |                                |
| AC-2          | VSTAA   | WPI               | S       | A       | T       | FWLIANSW--DWGE                 |
| HmCP1         | VSAAE   | MDH               | A       | A       | -       | YWNVANSW--DW--                 |
| HmCP2         | VSAAS   | YDH               | A       | A       | V       | YWTVANSW--DWGG                 |
| GCP-7         | VSAAQ   | YNA               | T       | S       | G       | YWLIANSW--DWGE                 |
| sc-0041161-1  | VSAAS   | YNG               | T       | S       | G       | YWTIANSW--DWGE                 |
| sc-0041161-2  | VSAAQ   | YNA               | T       | S       | G       | -                              |
| sc-0041161-3  | VSAAE   | -                 | -       | -       | -       | -                              |
| sc-0059702-4  | VAAAS   | SIF               | A       | A       | Q       | YWTVANSW--DWGE                 |
| sc-0059702-5  | VSAAE   | SLM               | A       | A       | S       | YWTVANSW--DWGE                 |
| sc-0059702-6  | VSAAE   | LAI               | A       | A       | Y       | YWTVANSW--DWGE                 |
| sc-0059702-7  | VSAAE   | AIR               | A       | A       | E       | YWTVANSW--YWGE                 |
| sc-0059702-8  | VSAAE   | YTI               | A       | A       | -       | YWTIANSW--DWGE                 |
| sc-0059702-9  | -   | -                 | A       | A       | E       | YWTVANSW--WGEE                 |
| sc-0059702-10 | VSAAE   | YAI               | A       | A       | E       | YWTIANSW--DWGE                 |
| sc-0059702-11 | VSAAS   | YSI               | A       | A       | D       | YWTIANSW--DWGE                 |

**Table 4.6 Conserved regions within *H. contortus* GCP-7 and the proteases present on supercontig\_0041161 and supercontig\_0059702**

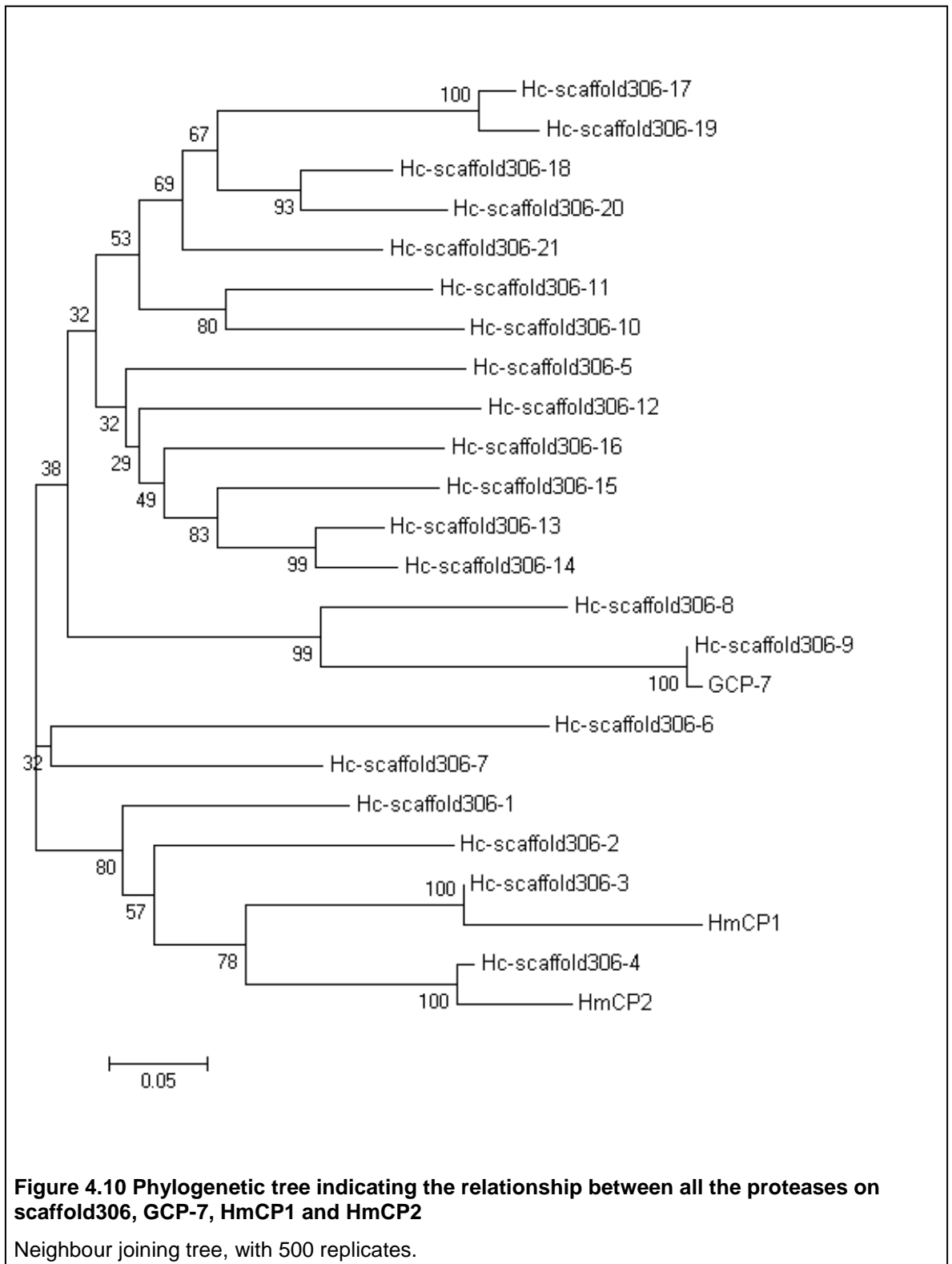
AC-2, HmCP1 and HmCP2 have been added to the table as a reference to the previously characterised AC protease family and those proteases identified as being similar to GCP-7.

The newly released *H. contortus* sequence data has been analysed in an attempt to complete the sequences for these GCP-7-related proteases. A BLASTn search against the haem\_supercontigs.12062012 file on the Sanger ftp site, using the complete sequence for both supercontig\_0041161 and supercontig\_0059702, identified that both are present on scaffold306. This scaffold is 186 kb in length and provided the full coding sequence for a number of the proteases that on the supercontigs were incomplete. Ten additional GCP-7-like proteases were also identified and for the majority of these the coding sequence is complete. Figure 4.9 is an Artemis screenshot of scaffold306 indicating the structure of the genes present, how the scaffold relates to the previously identified supercontig sequences and also the names of the new proteases. Analysis of the cysteinyl active site regions suggested that the previously identified HmCP1 and HmCP2 are more like GCP-7 than the other HmCP proteases. Interestingly, proteases with high percentage identity to HmCP1 (Hc-scaffold306-3, 93% AA identity, 98% DNA identity) and HmCP2 (Hc-scaffold306-4, 88% AA identity, 97% DNA identity) were identified on this scaffold, further supporting the hypothesis that GCP-7, HmCP1 and HmCP2 are closely related. Figure 4.10 is a phylogenetic tree indicating the relationship between all the proteases on scaffold306, GCP-7, HmCP1 and HmCP2. This figure indicates that the majority of the proteases present on this new scaffold and not previously identified on the supercontigs are more similar to HmCP1 and 2 than to GCP-7.



**Figure 4.9 Annotation of the protease genes on scaffold306 (12062012 file)**

The order displayed is the order of proteases on the scaffold. The scaffold sequence obtained from the ftp site has been reversed so that proteases are displayed on the positive strand in the forward orientation. The regions covered by supercontig\_0041161 and supercontig\_0059702 are indicated. There is only one exon present on the scaffold for the last protease therefore the protease sequence on supercontig\_0059702 has been included to complete the family. Of the 21 proteases identified, those highlighted in red have a high identity with previously published partial and complete gene sequences. The accession numbers for these sequences are in Appendix 3 Table 2.



The sequence information gained from the annotation of the two supercontigs and scaffold306 indicates that there are at least 21 GCP-7-like proteases in the *H. contortus* genome. In the study by Rehman and Jasmer (1998) only one GCP-7 sequence was identified. This is the first analysis to show GCP-7, like the AC and HmCP proteases, is a member of a multigene family of tandemly arranged

cathepsin B proteases. It also provides evidence that both HmCP1 and HmCP2 are more closely related to GCP-7 than to the other HmCPs.

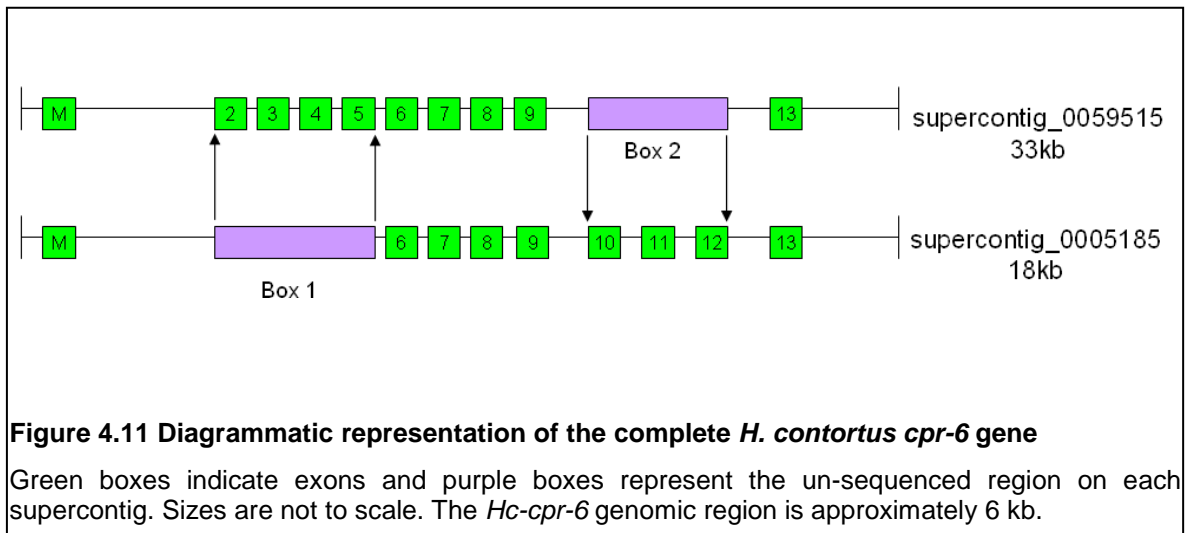
## 4.2.2 Analysis of *H. contortus* *cpr-6*, a unique protease gene

### 4.2.2.1 Completion and annotation of the conserved *H. contortus* cysteine protease gene *cpr-6*

Larminie and Johnstone (1996) identified a cathepsin B cysteine protease gene family in *C. elegans* containing six members named *cpr1-6*. However the *C. elegans* *cpr* genes are not highly similar, with the encoded proteases showing only 40-65% identity. Southern blotting, together with genome data confirmed that each of the six identified genes is single copy. RT-PCR showed that *cpr-6* was the only gene in which transcript levels increased abundantly during the L4 and adult stages, with a 17.9-fold difference between minimum and maximum levels. This *C. elegans* cysteine protease is highly conserved in *A. suum* and partial sequences identified in *H. contortus* (Rehman and Jasmer, 1999). This is in contrast to other cathepsin B cysteine proteases which are not highly conserved between *C. elegans* and parasitic nematodes. It was of interest to determine if the *Hc-cpr-6* gene is tandemly repeated and part of a multigene family. tBLASTn analysis was carried out using the *C. elegans* *cpr-6* sequence against the database of *H. contortus* assembled supercontigs (21/08/08)(all reads). This identified a region on supercontig\_0059515 which was annotated using Artemis. There are no other genes similar to *cpr-6* on this supercontig or elsewhere in the genome, confirming that, in contrast to all the other cathepsin B sequences, *H. contortus* *cpr-6* is a single copy gene. The complete *H. contortus* *cpr-6* gene could not be identified on this supercontig nor could any additional sequence be identified from the parasite EST database website <http://www.ebi.ac.uk/Tools/sss/wublast/parasites.html>.

Sequence information was missing at both the start and end of *H. contortus* *cpr-6* therefore 5' and 3' RACE were carried out on adult worm cDNA (Chapter 2.3.1.3). A proposed start methionine codon was identified, and similar to the *H. contortus* CBL cysteine protease genes annotated previously, it is located on a separate exon. The 3' end was identified, however due to sequencing gaps, alignment to supercontig\_0059515 was not possible. A BLAST search of the *H.*

*contortus cpr-6* 5' RACE sequence was carried out against the newer supercontig database on the Sanger website, *H. contortus* supercontigs (26/08/2009), and a high similarity match of this region was identified on supercontig\_0005815. Sequence alignments showed that this supercontig contained large stretches of identity to supercontig\_0059515. On both supercontigs the sequence at each side of the gaps corresponds to the other and from this it can be assumed that these supercontigs cover the same genomic region and provide the complete *H. contortus cpr-6* sequence (Figure 4.11). BLAST analysis of the 12062012 sequence file also confirmed this gene annotation, and that *cpr-6* is a unique gene within *H. contortus*.



#### 4.2.2.2 CPR-6 conservation in other parasitic nematodes

Due to the high level of conservation of *cpr-6* in *H. contortus*, *C. elegans* and *A. suum*, other parasitic nematode genomic and EST sequences were analysed. This aimed to determine whether CPR-6 is highly conserved over a wider range of nematodes and what the putative function of CPR-6 may be. CPR-6 related sequences for a number of nematodes were identified; *C. elegans* from wormbase, *H. contortus*, *Nippostrongylus brasiliensis*, *Teladorsagia circumcincta* and *Globodera pallida* from the Sanger Blast server, *A. suum* and *Brugia malayi* from NCBI and *Heligmosomoides polygyrus*, *Dirofilaria immitis*, and *Litomosoides sigmodontis* from the 959 Nematode Genome Blast Server <http://xyala.cap.ed.ac.uk/downloads/959nematodegenomes/blast/blast.php>.

The percentage identity of CPR-6 across the different nematode species was calculated and the results shown in Table 4.7. Unsurprisingly CPR-6 from the clade V nematodes (*H. contortus*, *T. circumcincta*, *N. brasiliensis*, *H. polygyrus* and *C. elegans*) all share a higher percentage identity with each other than nematodes in different clades. Interestingly, *A. suum* (clade III) CPR-6 shows the highest similarity to clade V rather than other clade III nematodes (*L. sigmodontis*, *B. malayi* and *D. immitis*).

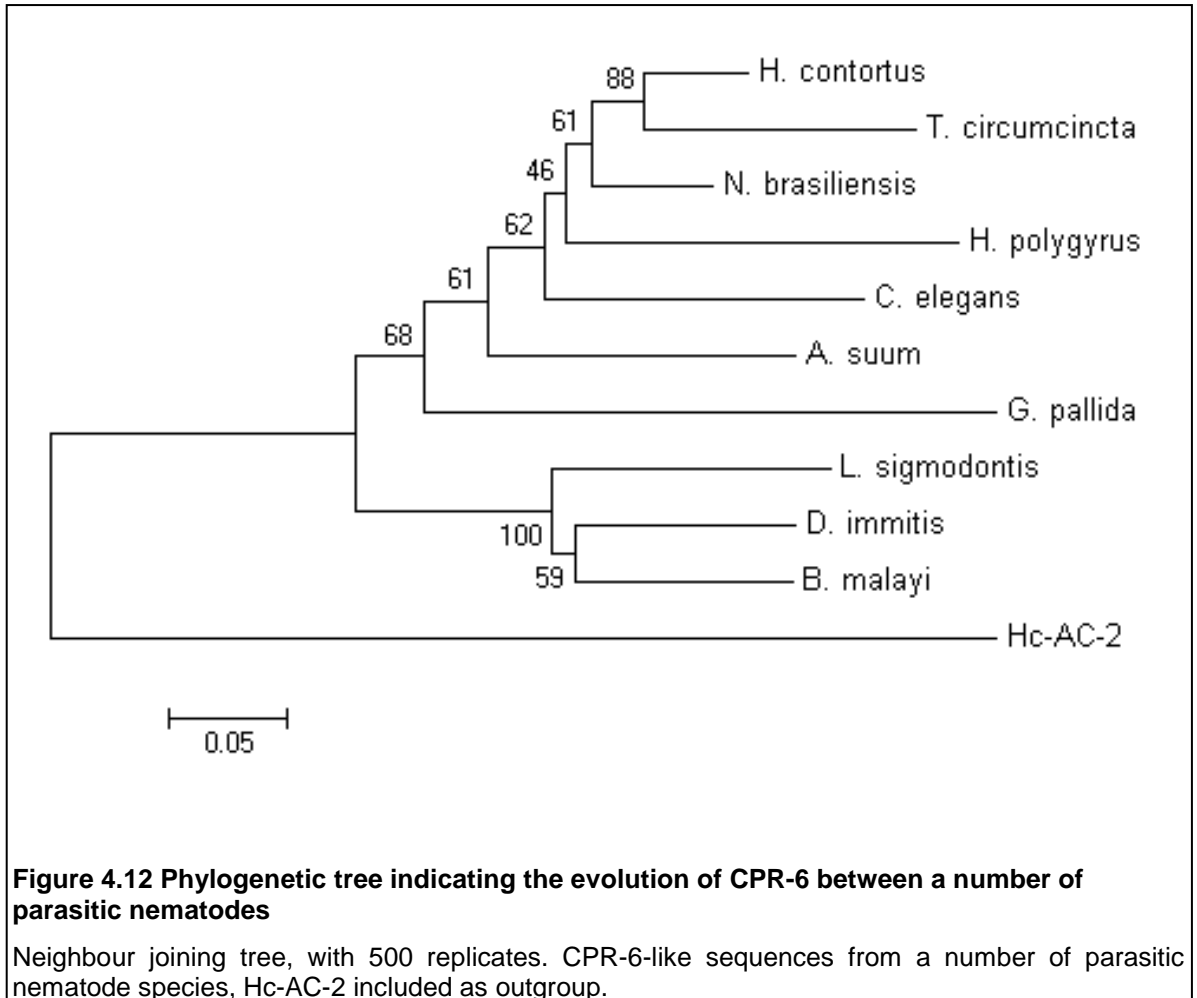


| Species                | <i>H. contortus</i> | <i>C. elegans</i> | <i>A. suum</i> | <i>N. brasiliensis</i> | <i>T. circumcincta</i> | <i>H. polygyrus</i> | <i>D. immitis</i> | <i>B. malayi</i> | <i>L. sigmodontis</i> | <i>G. pallida</i> |
|------------------------|---------------------|-------------------|----------------|------------------------|------------------------|---------------------|-------------------|------------------|-----------------------|-------------------|
| <i>H. contortus</i>    | 100                 | 70                | 68             | 83                     | 82                     | 78                  | 63                | 58               | 61                    | 61                |
| <i>C. elegans</i>      |                     | 100               | 65             | 71                     | 66                     | 73                  | 57                | 55               | 56                    | 61                |
| <i>A. suum</i>         |                     |                   | 100            | 67                     | 66                     | 69                  | 62                | 62               | 59                    | 62                |
| <i>N. brasiliensis</i> |                     |                   |                | 100                    | 78                     | 79                  | 63                | 59               | 60                    | 61                |
| <i>T. circumcincta</i> |                     |                   |                |                        | 100                    | 64                  | 57                | 57               | 55                    | 53                |
| <i>H. polygyrus</i>    |                     |                   |                |                        |                        | 100                 | 60                | 58               | 58                    | 59                |
| <i>D. immitis</i>      |                     |                   |                |                        |                        |                     | 100               | 72               | 69                    | 55                |
| <i>B. malayi</i>       |                     |                   |                |                        |                        |                     |                   | 100              | 70                    | 56                |
| <i>L. sigmodontis</i>  |                     |                   |                |                        |                        |                     |                   |                  | 100                   | 55                |
| <i>G. pallida</i>      |                     |                   |                |                        |                        |                     |                   |                  |                       | 100               |

**Table 4.7 Percentage identity at the amino acid level of CPR-6 in a number of parasitic species**

Species with the highest percentage identities have been highlighted in red.

Phylogenetic analysis was carried out to identify the relationship between the cysteine proteases (Figure 4.12). As expected from the percentage identity, those species in the same clade are located closer together in the tree. The clade V nematodes are grouped together with the addition of *A. suum*, which also groups with this clade rather than the other clade III nematodes. The remaining clade III nematodes group together as expected.



The CPR-6 proteases were aligned using Clustal W2 and the results indicated in Figure 4.13 with active site regions highlighted. As mentioned previously CBL cysteine proteases contain a number of important conserved regions within their structure (Figure 4.13 and Table 4.8). In contrast to all the other *H. contortus* cathepsin B proteases identified, CPR-6 proteases can be characterised as Type A (FGAVE) (Rehman and Jasmer, 1999). The three filarial clade III nematodes have a slightly different sequence from the others (VAAVE). The S2 subsites are also shown in Table 4.8, and again CPR-6 sequences from clade III nematodes are more divergent.

The UniProt website (<http://www.uniprot.org/uniprot/P43510>) was used to identify structural characteristics of *C. elegans* CPR-6. This website provided information about the signal peptide region and cleavage site, and indicated that there is a pro-peptide region, from amino acid 17-104, resulting in a mature protein from amino acid 105-379. As the CPR-6 cysteine proteases show strong conservation at the amino acid level it would be expected that known structural features identified in *C. elegans* will be very similar in the other species analysed. Six disulphide bonds have been identified in the *C. elegans* CPR-6 sequence (Figure 4.13) and the cysteines involved are conserved in all nematode CPR-6 proteases identified here, suggesting conserved folding.

|                |  |     |
|----------------|--|-----|
| D.immitis      | -----ISGLLICIFHITKR-----RIEQKSLKNVKKNVSEGRDDQ----DEYKNKKIS       | 45  |
| B.malayi       | MYLLPLLWFFLLPLYLTAVY-----ALVTDPESSTKSLWQSSKEKKRFVIDKYRNTKIA      | 55  |
| L.sigmodontis  | -----ICITTFASSHF-----DRKIEMDENEWERIWERKHFQGRKLDEYRNIETG          | 46  |
| H.contortus    | -----MKGLLVCLFFVLADR-----RIELDSSSEYQKEGRDFQNRSTLESFRNRIP         | 49  |
| T.circumcincta | -----MRGFLVCLLFVLVEC-----RIKIDLSSSEEEYPE-KDFTNRRSTLESFRNRKIP     | 48  |
| N.brasiliensis | -----MRGYLVCLLFALAQC-----RVELYSSSEYDVKPQR-QLRRSTLESFRNRKIP       | 48  |
| H.polygyrus    | -----IP  | 2   |
| C.elegans      | -----MKTLLFLSCLVVAAY-----CACN-----DNLESVLDKYRNRID                | 35  |
| A.suum         | -----MRYTFLVALLAIANSSADRRLHIDDSSSSDESYSGRYYGDRSVLDEFDRDKIS       | 54  |
| G.pallida      | -----DSSSESETSPARGAYTRVSIAGELRTARPS                              | 33  |
|                |  |     |
| D.immitis      | SNARNLSGQELIDYINSYQTLWKAENVK-FNLYS--DEVKYGLMGNVNLKQLLTNKKK       | 100 |
| B.malayi       | PEAENLSGQELIDYVNSHQTLWKAAGMNK-FNLYS--DTVKYGLLGNVNRKKSVEHKKN      | 110 |
| L.sigmodontis  | PEARNLSGQELIDYVNSRQTLWKAAGNTK-FNSYD--ATVKYGLLGVSDLE-----KKK      | 96  |
| H.contortus    | TEAEELTGRELIDYVNSHQSLWRAKENRRFARYP--DRTKWGLMGNVNVRLSVRAKQH       | 105 |
| T.circumcincta | TEAEELTGKELIDYVNSHQSLWKAENRRFARYP--DRTKWGLMGNVNVRLSVMAKQH        | 104 |
| N.brasiliensis | AEAEQLTGRELIDYVNSKQSLWKAENRRFKHYP--DHTKWGLMGNVNVRLSVKAKQH        | 104 |
| H.polygyrus    | SEAE-LTGKDLIDYVNRKQSLWKAKEHRRFSRYP--DRTKWGLMGNVNVRLSVKAKQH       | 57  |
| C.elegans      | SEAAELDGDDLIDYVNNQNLTAKKQRRFSSVYGENDKAKWGLMGNVNVRLSVKAKQH        | 94  |
| A.suum         | HEAEKLTGYALANYVNRKQNLWKAENRRFKHYP--DRVKYGLMGNVNVRLSVKAKKN        | 110 |
| G.pallida      | PRAEQLRGQSLVDYVNGRQGLWRAELSPKFESYD--ESVKWRMMGNVNVRLSVKAKKM       | 89  |
|                |  |     |
|                | I * * *  |     |
| D.immitis      | LSPTRHFVHVPESFDAREKWPCEASLRNIRDQSSCGSCWAVAAVEAMSDRICIMSKGK       | 159 |
| B.malayi       | LSPIRHSNIFIPESFDARKNWPCEASLRNIRDQSSCGSCWAVAAVEAMSDRICITSKGK      | 169 |
| L.sigmodontis  | LPVMQYSNSYIPESFDAREKWPCEASLRNIRDQSSCGSCWAVAAVEAMSDRICIMSKGK      | 155 |
| H.contortus    | LSTTKDLLDIDIPESFDAREEWPDCSISIKVIRDQSSCGSCWAFGAVEAMSDRICIASKGE    | 164 |
| T.circumcincta | LSTTKDLLDIDIPESFDSREEWPNCEISIKVIRDQSSCGSCWAFGAVEAMSDRICIASKGE    | 163 |
| N.brasiliensis | LSATKDLLDIDIPESFDSREQWPECSISIKVIRDQSSCGSCWAFGAVEAMSDRICIASKGE    | 163 |
| H.polygyrus    | LSATKDLLDIDIPESFDSREQWPECSISIKNIRDQSSCGSCWAFGAAEAITDRICIESKGS    | 116 |
| C.elegans      | LSKTKDLLDIDIPESFDSRDNWPCKDSIKVIRDQSSCGSCWAFGAVEAMSDRICIASHGE     | 153 |
| A.suum         | LSPTRFYDIYIPEAFDAREKWDQCASLRNIRDQSSCGSCWAFGAVEAMSDRICIASNGK      | 169 |
| G.pallida      | LGKTRFLDLLDLPDSFDARQQWPFPSISGLIRDQSSCGSCWAFGAVEAMSDRTCIASGQK     | 148 |
|                |  |     |
|                | ** * *   |     |
| D.immitis      | IQVTL SADDLLSCCKTCGFGCYGGNPIEAWKYWVSDGIVTGSNYTNHSGCRPYPFPPCEH    | 219 |
| B.malayi       | KQVIL SADDLLSCCKTCGFGCFGGEPMAAWKYWVLSGIVTGSNTYTNHSGCRPYPFPPCEH   | 229 |
| L.sigmodontis  | VQVTL SADDLLSCCRCTCGFGCYGGIPVAAWKYWISSGIVTGSNTNHTGCRPYPFPPCEH    | 215 |
| H.contortus    | IQVLS ADDLLSCCKSCFGGCGNGGDLPLAAWKYWVKDGIVTGSNFTANQGCKPYPFPPCEH   | 224 |
| T.circumcincta | IQVLS ADDLLSCCKSCFGGCGNGGDLPLAAWKYWVRDGI V TGSNFTANQGCKPYPFPPCEH | 223 |
| N.brasiliensis | IQVLS ADDLLSCCKSCFGGCGNGGDLPLAAWKYWVKSGIVTGSNFTMNEGCKPYPFPPCEH   | 223 |
| H.polygyrus    | FKPEISADELLACCDTCGEGCGNGGDLPLSAWKYWVKDGIVTGSNFTANQGCKPYPIPPCGH   | 176 |
| C.elegans      | LQVTL SADDLLSCCKSCFGGCGNGGDLPLAAWRKYWVKDGIVTGSNYTANNGCKPYPFPPCEH | 213 |
| A.suum         | IQVLS ADDLLSCCKSCFGGCDGGEPMAAWKYWVKEGIVTGSNFTMKQGCKPYPFPPCEH     | 229 |
| G.pallida      | IQITL SADDLLSCCRKCGFGCDGGEPLQAWRFWVKEGIVSGSNFSVHGGRPYPFPPCEH     | 208 |

|                |  |   |  |
|----------------|--|---|--|
|                |  | *                    *                    * |  |
| D.immitis      | HNKTHYKQCRHDLPTPKCYKCKRKYDYG-KSYEADKYYGKQAYSVGNDESIOKEIMTM   | 278   |  |
| B.malayi       | HSNKTHYEPCKHDLPTPKCYKQCDKNYT-KSYKADKYYGEQAYNVENDVESIOKEIMTL  | 288   |  |
| L.sigmodontis  | HSNKTHYKPCRHDLPTPKCYKCKRKYDYG-KSYRSDKYYGKAYGVDSDEVAIQREIMTM  | 274   |  |
| H.contortus    | HSNKTHYDPCRHLDFPTPKCEKRCVPTYNEKTYNDDKYYGRNAYGVKDDVTAIQKEVLTH | 284   |  |
| T.circumcincta | HSNKTHYDPCRHLDFPTPKCEKRCVPTYNEKS---DEFSGATAYAVSKKVTDIQKEIMTN | 280   |  |
| N.brasiliensis | HSNKTHYDPCRHLPTPKCEKTCVPSYKGRSYTEDKYYGRSAYGVKDDVTAIQKEIMTH   | 283   |  |
| H.polygyrus    | HANETYFGPCPTDEYDTPVCTKKCIAGYA-TAYADDKHYGKNAYGVKDDVTAIQKEILTY | 235   |  |
| C.elegans      | HSKKTDFDPCPHDLPTPKCEKCKVSDYTDKTYSEDKFFGASAYGVKDDVEAIQKELMTH  | 273   |  |
| A.suum         | HSNKTHYQPCRHDLPTPKCEKRCVPTYNEKTYAEDKFFGETAYGVEDDVTSIQKEILTH  | 289   |  |
| G.pallida      | HSNKTHFEPCKTELYPTPKCEKRCVPSYKGRSYTEDKYYGRSAYAVENSMKAIQNELVYN | 268   |  |
|                |  |   |  |
| D.immitis      | GPVEVAFEVHTDFLNYAGGIYKHVAGSMIGGHAVKMLGWGIDQGVPTWLAANGWNTDWC  | 338   |  |
| B.malayi       | GPVEVAFEVYDFLHYTSGIYKHVAGSVGGGHAVKILGWGIDQGVPTWLAANGWNTDWC   | 348   |  |
| L.sigmodontis  | GPVEVAFEVYDFLQYTGGIYKHLAGSVGGGHAVKILGWGIDQGVPTWLAANGWNTDWC   | 334   |  |
| H.contortus    | GPVEVAFEVYDFLNYAGGIYVHTGGRLGGGHAVKILGWGVEQGMPTWLIANGWNTDWC   | 344   |  |
| T.circumcincta | GPVEVAFEVYDFEHYTGGIYVHTWGAESGGHAVKVIWGTETGPTWLVANGWNTDWC     | 340   |  |
| N.brasiliensis | GPVEVAFEVYDFLNYAGGIYVHTGGKLGGHAVKMIWGVQGMPTWLVANGWNTDWC      | 344   |  |
| H.polygyrus    | GPVEVAFEVYDFLNYAGGIYVHTGGKLGGHAVKMIWGVQGMPTWLVANGWNTDWC      | 295   |  |
| C.elegans      | GPLEIAFEVYDFLNYDGGIYVHTGGKLGGHAVKLIWGIIDGIPYWTVANWNTDWC      | 333   |  |
| A.suum         | GPVEVAFEVYDFLNYDGGIYVHTGGKLGGHAVKMLGWGVEQGMPTWLVANGWNTDWC    | 349   |  |
| G.pallida      | GPVEVAFEVYDFMNYKGGVYVHTGGKLGGHAVKLLGWADNGIPTWLVANGWNTDWC     | 328   |  |
|                |  |   |  |
| D.immitis      | D---GYFRILRGVDECGVESGIVGGIPK-----                            | 363   |  |
| B.malayi       | DVFSGYFRILRGADECGIESGIVAGIPRKDARSKAR-----                    | 384   |  |
| L.sigmodontis  | D---GYFRIRGIDECEGIESGIVAGIPKRRGKSKFH-----                    | 367   |  |
| H.contortus    | D---GYSRILRGVDECGIESGVVGGVPKLNISIHRRRHVAASSYTDFF----         | 389   |  |
| T.circumcincta | K---GF-RI-----   | 345   |  |
| N.brasiliensis | D---GYFRILRGVDECGIESGVVGGIPKINSIHRRRRHHR-WYTDD-----          | 385   |  |
| H.polygyrus    | D---GFFRILRGVDECGIESGVVGGVPKINSI-----                        | 324   |  |
| C.elegans      | D---GFFRILRGVDECGIESGVVGGIPKLNLSLTSRLHRRHRRHVYDDNY---        | 379   |  |
| A.suum         | D---GFFRIIRGIDECEGIESGVVGGLPKLNRTYKRYHRRYRLDNDEDDDIIF        | 398   |  |
| G.pallida      | D---GFFRILRGKDECGI-----                                      | 343   |  |

**Figure 4.13 Amino acid alignment of CPR-6 from *H. contortus* and other parasitic nematodes**

The signal peptide region is highlighted in yellow, the three active site regions in orange and the S2 subsites in blue. The occluding loop region is highlighted in grey and the haemoglobinase motif indicated in green. The pro-peptide cleavage site is indicated by I. All cysteine pairs, potentially linked by a disulphide bond are indicated (star). Highlighted in brown are the peptide regions described in Section 4.2.2.5.

| Sequence               | Cysteiny active site signature sequence (147-151) | S2 Subsite        |         |         |         | Haemoglobinase motif (331-344) |
|------------------------|---|-------------------|---------|---------|---------|--------------------------------|
|                        |   | Y, P, S (189-191) | A (290) | A (317) | E (362) |                                |
| <i>H. contortus</i>    | FGAVE   | DPL               | A       | A       | G       | YWLIANSW--DWGE                 |
| <i>C. elegans</i>      | FGAVE   | DPL               | A       | A       | G       | YWTVANSW--DWGE                 |
| <i>A. suum</i>         | FGAVE   | DPM               | A       | A       | S       | * YWLVANSW--DWGE               |
| <i>N. brasiliensis</i> | FGAVE   | DPL               | A       | A       | G       | * YWLVANSW--DWGE               |
| <i>T. circumcincta</i> | FGAVE   | DPL               | S       | A       | -       | * YWLVANSW--DWGE               |
| <i>H. polygyrus</i>    | FGAVE   | DPL               | A       | A       | G       | * YWLVANSW--DWGE               |
| <i>D. immitis</i>      | VAAVE   | NPI               | S       | A       | G       | YWLAANSW--DWGE                 |
| <i>B. malayi</i>       | VAAVE   | EPM               | S       | A       | G       | YWLAANSW--DWGE                 |
| <i>L. sigmodontis</i>  | VAAVE   | IPV               | S       | A       | G       | FWLAANSW--DWGE                 |
| <i>G. pallida</i>      | FGAVE   | EPL               | A       | A       | -       | YWLVANSW--DWGE                 |

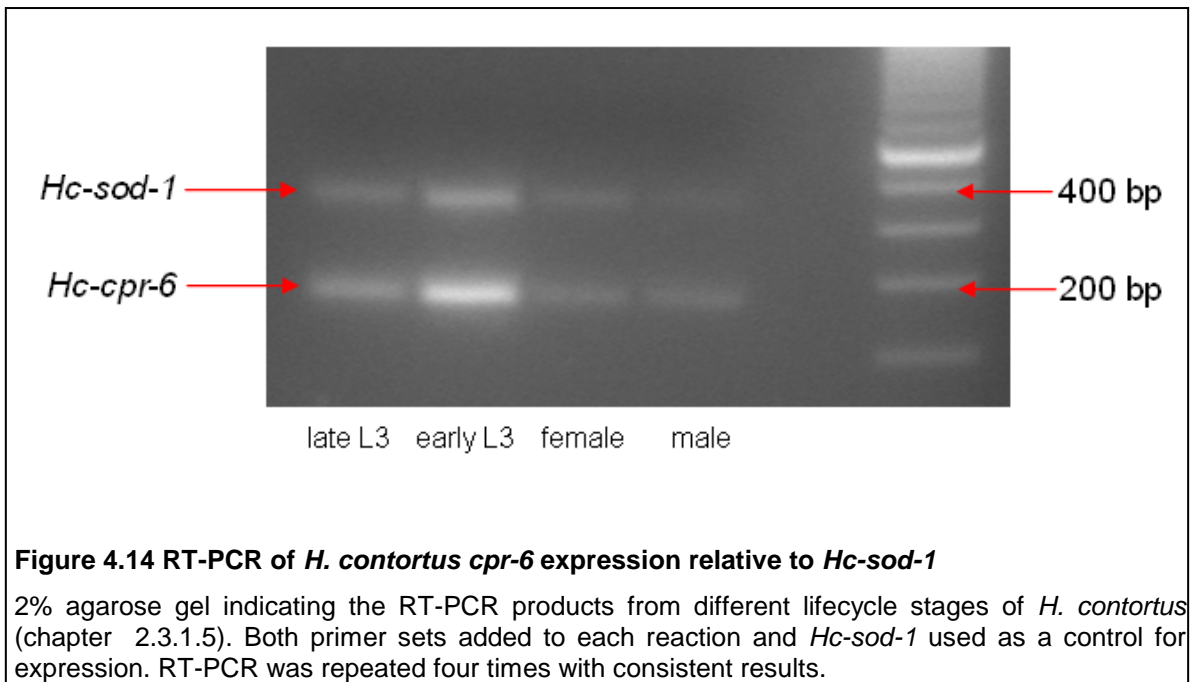
**Table 4.8 Conserved regions within the CPR-6 protease in a number of nematodes**

A star indicates the parasites in which the haemoglobinase motif is conserved with the consensus. Dashed lines indicate where sequence is unavailable.

#### 4.2.2.3 *cpr-6* expression in *H. contortus* adult and larval stages

BLAST analysis was carried out against Nembase 4 ([www.nematodes.org/nembase4/](http://www.nematodes.org/nembase4/)) to identify ESTs for *H. contortus cpr-6*. 95% sequence identity was used as a cut off when identifying potential matches and four ESTs (adult stage) in total were found using these criteria. Cluster number HCC01809\_1 has a 97% sequence identity (663/679 bp) and has three associated ESTs. Cluster number HCC09542\_1 has a 98% sequence similarity (433/439 bp) and has one associated EST. This information indicates that *H. contortus cpr-6* is expressed, albeit at a lower level than other *H. contortus* cathepsin B genes.

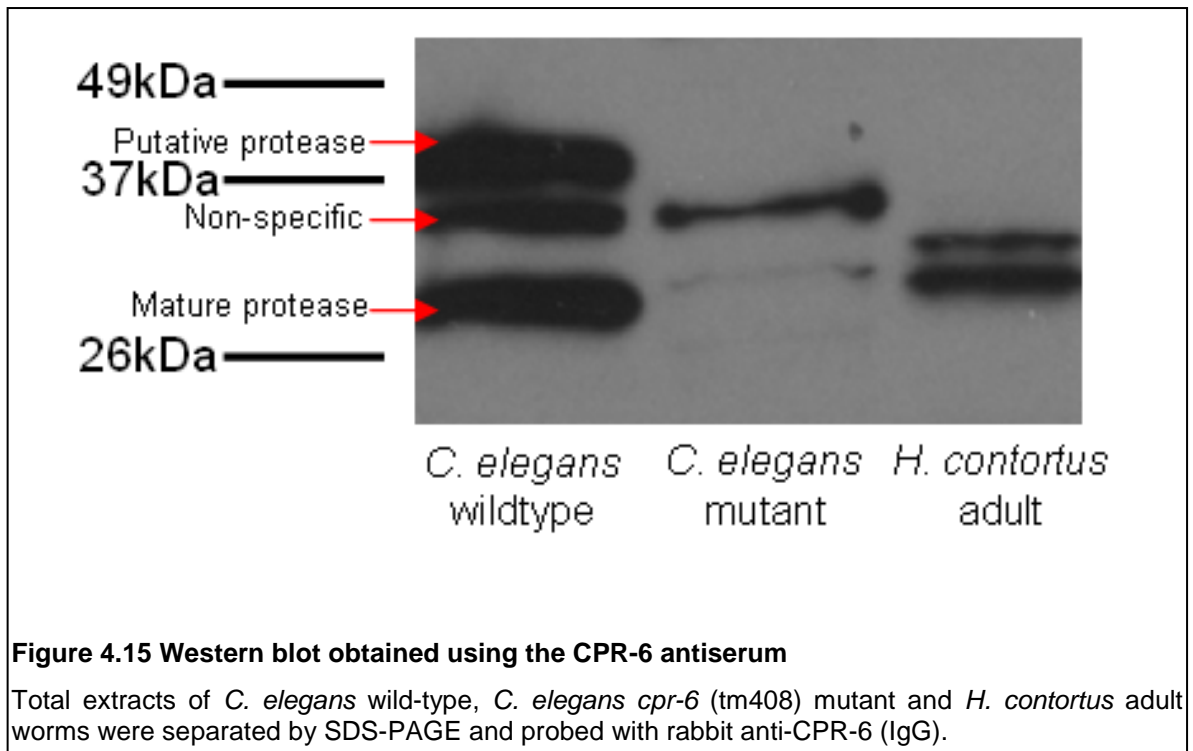
Previous analysis has shown that *H. contortus* cathepsin B genes (HmCP family) are expressed only in L4 and adult stages, suggesting a role in feeding (Skuce *et al.*, 1999). Given the strong conservation of CPR-6 across a range of nematode species it was of interest to examine the temporal expression of the *Hc-cpr-6* gene. *cpr-6* expression levels were examined in adult and larval stages of *H. contortus* worms, relative to constitutively expressed *Hc-sod-1* (Liddell and Knox, 1998). Male, female (both 28 days post infection), early L3 (24 hour cultured) and late L3 (72 hour cultured) cDNA was used to test expression (Chapter 2.3.1.3). RT-PCR was carried out using gene specific primers designed to give a product size of 271 bp for *cpr-6* and 403 bp for *Hc-sod-1* (primers in Appendix 2 Table 4.1) (Figure 4.14). This semi-quantitative RT-PCR approach indicated that *Hc-cpr-6* is slightly more highly expressed in the L3 stage than in day 28 adult worms, albeit expression level is quite low.



#### 4.2.2.4 Attempts to localise CPR-6 using specific antibody

To examine localisation of CPR-6 in *H. contortus* and *C. elegans*, CPR-6 antibodies were generated. Two peptide regions highly similar in *H. contortus* and *C. elegans* CPR-6 were identified: **SFDSRDNWPKCDSIKV** and **PHDLYPTPKCEKKCV** (Figure 4.13). These two sequences were used to generate antibody in rabbits (Chapter 2.1.6.1).

The antiserum generated was tested by western blotting to confirm recognition of CPR-6, using extracts of *C. elegans* N2 wild-type, a *C. elegans* *cpr-6* mutant strain (allele tm408) and *H. contortus* adult worms. The *C. elegans* CPR-6 mutant contains a 310 bp deletion in the *cpr-6* coding sequence. In both *C. elegans* and *H. contortus* the CPR-6 pro-enzyme is predicted to be 43 kDa and the mature protease 30 kDa. Bands of the expected sizes were observed for *C. elegans* wild-type worms. An additional protein of approximately 35 kDa was detected which is most likely due to non-specific reactivity and could be observed in the mutant extract. A band of approximately 32 kDa was detected in the *H. contortus* adult extract which may represent the mature protease (Figure 4.15). However, despite reactivity by western blotting, the CPR-6 antibody showed no specific binding to *C. elegans* or *H. contortus* by immunofluorescence, therefore an alternative approach was taken.



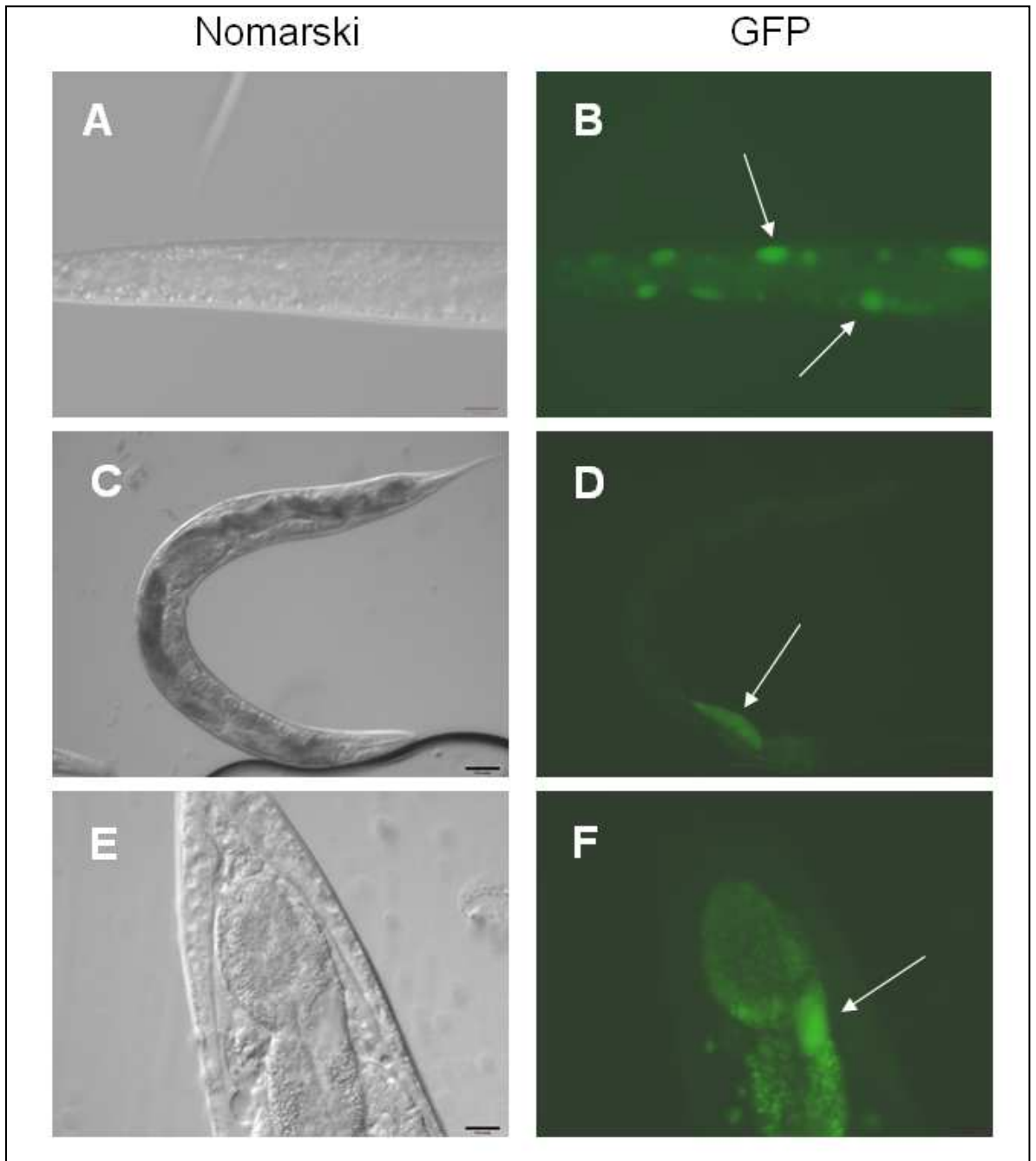
**Figure 4.15 Western blot obtained using the CPR-6 antiserum**

Total extracts of *C. elegans* wild-type, *C. elegans cpr-6* (tm408) mutant and *H. contortus* adult worms were separated by SDS-PAGE and probed with rabbit anti-CPR-6 (IgG).

#### 4.2.2.5 Temporal and spatial expression of a *C. elegans cpr-6* translational reporter

Previous studies have demonstrated that temporal and spatial expression driven by parasite gene promoters can be identified in transgenic *C. elegans* (Britton *et al.*, 1999). Expression of a *Hc-cpr-6::gfp::lac-Z* transcription fusion in *C. elegans* was attempted (primers in Appendix 2 Table 4.2). However, using a 1.2 kb upstream region of *Hc-cpr-6* fused to GFP/Lac-Z failed to show any detectable expression. As *cpr-6* is not highly expressed in *H. contortus*, it is possible that the promoter is not highly active. Analysis of *Ce-cpr-6* was therefore carried out. A PCR product was amplified containing 1.7 kb promoter sequence and the complete *Ce-cpr-6* gene sequence with the omission of the stop codon (primers in Appendix 2 Table 4.3). A transgenic line containing this translational fusion construct was obtained and GFP expression was visible under the bench microscope in adult worms, and in both adults and larvae at high power. *Ce-cpr-6::gfp* localised to intestinal cells and was present down the entire length of the worm (Figure 4.16). Interestingly, GFP expression increased in older worms (>10 day old adults) and in some dauer stage larvae.





**Figure 4.16 *Ce-cpr-6::gfp* translational fusion in transgenic *C. elegans* worms**

(A and B) Young adult stage showing GFP fluorescence foci in hypodermal and gut basement membrane regions, image at x40 magnification. (C-E) Older adult worms showing GFP accumulation in anterior gut. C and D at x10 magnification and E and F at x40 magnification. Arrows indicate GFP fluorescence, additional fluorescence in panel F is autofluorescence. The 3.3 kb *Ce-cpr-6* gene sequence was cloned into GFP Fire vector pPD95.75 and injected into *C. elegans* wild-type hermaphrodites (10 ng/ $\mu$ l).

### 4.3 Discussion

The aim of this chapter was to examine the organisation and expansion of other cathepsin B multigene families from the *H. contortus* genome data. Novel genes related to the *H. contortus* cysteine protease HmCPs and the *gcp-7* gene were identified. In addition to this *Hc-cpr-6* was identified as a single copy gene which is found in a number of parasitic species and the level of conservation was analysed.

Work by Skuce *et al.*, (1999) identified the six member HmCP family; in addition to this a gut cysteine protease was identified by Rehman and Jasmer (1998) and termed GCP-7. It was unknown whether this seventh gut cysteine protease gene was an additional member of this protease family. Analysis contained in this chapter identified that HmCP1 and HmCP2 share only a 36-42% identity at the amino acid level with the other HmCPs within the family and also those proteases identified on supercontig\_0059492 (for which a significant amount of sequence data was available). In addition, phylogenetic analysis grouped HmCP1 and HmCP2 separately and the cysteinyl active site signature sequence also differed from others within this family. On analysis of GCP-7 and the GCP-7-like proteases identified on supercontig\_0041161 and supercontig\_0059702 it became apparent that HmCP1 and HmCP2 are more similar to these proteases with 47-64% amino acid identity. Thus HmCP1 and HmCP2 are most likely to be part of a multigene protease family with GCP-7 and the GCP-7 related proteases and not the six member HmCP family identified previously. Therefore, in addition to the 8 member tandemly arranged AC cysteine protease gene family identified in Chapter 3, there appears to be a CBL cysteine protease multigene family related to the HmCPs, containing 12 members, and a family of 21 tandemly arranged genes that are related to *gcp-7*. 41 individual genes have been identified here and it is possible that more cathepsin B genes are present in un-sequenced regions.

Rispe *et al.*, (2008) carried out a study of CBL cysteine protease genes in aphids and identified approximately 28 gene copies in the pea aphid *Acyrtosiphon pisum*. This finding was interesting as aphids feed on plant phloem which is high in sucrose and it was thought that sucrose hydrolysis occurred in place of food

digestion, thus it may be expected that there would be no requirement for proteases. In addition to genes with expression supported by EST data, there were CBL gene copies in the *A. pisum* genome that had no associated EST data. This indicates that they may be pseudogenes and on closer examination of the gene structure a stop codon was identified in one *A. pisum* CBL gene. As a number of the cysteine proteases identified in *H. contortus* had no associated ESTs it was of interest to examine the coding sequence for any stop codons or mutations resulting in the protein being unstable. While no stop codons were identified within any of the proteases, for three *gcp-7*-like genes on supercontig\_0059702 with no EST data, the two in which the second exon could be identified have no start methionine codon upstream with the correct splice sequence after it. This may be a potential explanation for the lack of expression data.

CBL genes have been identified as being members of multigene families in other parasitic helminths. Work carried out by Harrop *et al.*, (1995) identified two CBL cysteine proteases Ac-CP-1 and Ac-CP-2 in *Ancylostoma caninum*, and found that Ac-CP-1 has a 61% identity with *H. contortus* AC-1. Therefore like *H. contortus*, *A. caninum* contains more than one CBL cysteine protease gene. Other species in which CBL gene families have been identified include *Fasciola hepatica*, FCP1-7 CBL cysteine proteases (Heussler and Dobbela, 1994) and *Schistosoma mansoni*, Sm31 and Sm32 proteins (Klinkert *et al.*, 1989). In the protozoan parasite *Leishmania major* >20 CBL protease genes have been identified (Mottram *et al.*, 1998). In contrast, species such as the rat, bovine and human have only one CBL cysteine protease gene. As more genome data becomes available, particularly for the other clade V nematodes, it will be interesting to determine if CBL gene expansion occurs more widely.

In addition to potentially allowing high level or co-ordinated expression, changes in the active site regions could result in multiple CBL activities/specificities being expressed. Rehman and Jasmer (1999) identified that for the twenty cysteine protease genes analysed over a number of species, not one had conservation of the S2 subsite glu245. From analysis of all the *H. contortus* CBL cysteine proteases studied in this chapter and the previous chapter, only HmCP6 and six of the proteases on scaffold306 (GCP-7-like) have a glutamine in this position. It has been speculated that glu245 leads to cathepsin L, rather than

cathepsin B type activity (Rehman and Jasmer, 1999), consistent with cathepsin L activity associated with gut extracts.

Characterisation of the CBL cysteine protease gene families in *H. contortus* is of interest due to their importance as potential vaccine candidates and it is possible that the large degree of diversity may interfere with vaccine efficacy (Jasmer *et al.*, 2004). Previous work by Redmond and Knox (2004) identified that a 38% reduction in *H. contortus* worm burden could be achieved after vaccination with the mature forms of purified *HmCP-1, 4 and 6* expressed in bacteria, however no concurrent decrease in FEC was observed. A subsequent vaccine trial identified a 29 and 27% decrease in worm burden and FEC respectively (Redmond and Knox, 2006). From the gene family classification work carried out in this chapter, it has become apparent that the three HmCP proteases used in these trials may be members of different families and thus may have slightly different functions. It is possible that targeting proteases that are in the same family and more highly conserved, may enhance vaccine outcome. Also, the use of a higher number of family members may also produce higher degrees of protection, as at this point it is unknown which members of the family are the most active and/or most highly expressed.

Relevant to this is CPR-6 which is a single copy protease of *H. contortus* and is also highly conserved in a range of nematodes, suggesting evolutionary pressure for sequence conservation. Although *cpr-6* mutants of *C. elegans* are superficially wild-type, it was observed that on long-term storage, the *cpr-6* mutant allele tm408, produced very few dauer larvae compared to wild-type *C. elegans*. In addition, expression of a *cpr-6::gfp* fusion construct increases in some dauer larvae and in ageing worms. Additional experiments are planned to examine in more detail the functional role(s) of nematode CPR-6, in particular a potential role in autophagy which is important for tissue re-modelling during dauer development and in ageing (Melendez *et al.*, 2003).

In conclusion, bioinformatic annotation has greatly increased the number of family members for both the HmCP and GCP-7-like CBL proteases of *H. contortus* and shown that these are tandemly arranged on the genome, most likely due to recent gene duplication. This has provided evidence for at least 42 individual CBL genes within the *H. contortus* ISE strain. In contrast, *cpr-6* was found to be

present as a gene with no expansion. This is the only *H. contortus* CBL gene to be highly conserved across a range of nematode species, suggesting an important role. Further functional analysis of the *H. contortus* CBL proteases and vaccine studies targeting specific families may help in defining which proteases are important for parasite survival.

## **Chapter 5**

Regulation of gut gene expression in nematodes

## 5.1 Introduction

The ever expanding genomic information available for parasites is enabling in depth analysis of genes and their promoters. Many studies have been carried out looking at the promoter regions of parasite genes and their potential role in gene regulation (Britton *et al.*, 1999; Pillai *et al.*, 2005; Zhao *et al.*, 2007). In previous chapters *H. contortus* gut genes have been identified and characterised. Gut genes are relevant for analysis due to the role of the gut and associated enzymes in blood digestion. Thus identification of factors that control expression of these genes may lead to novel mechanisms to interfere with gut gene expression in general, rather than attempting to target individual gut enzymes. Hawkins and McGhee (1995) identified that a number of Transcription Factors (TFs) are important in the regulation of gut development and gene expression in *C. elegans*.

Promoter analysis combined with deletion/mutation studies have helped identify regulatory motifs in a number of *C. elegans* genes (Okkema *et al.*, 1993). Work by Egan *et al.*, (1995) identified that a 36 bp region of DNA at the 5' end of the *ges-1* gene, containing two GATA sequences, results in a loss of gene expression when deleted. Additionally when only one GATA sequence is removed, expression is lost in the posterior but not the anterior gut. This paper concluded that GATA sequences present in either orientation within the promoter region are involved in gene expression. It also suggests that different motifs or regions within a promoter can be responsible for gene expression in the anterior and posterior gut. It is thought that this may reflect a gradient of TF activity within the gut. Additionally, if there is a loss of the region directly upstream or downstream of these GATA sequences a decrease in expression in the posterior gut cells is also observed. Subsequent studies showed that ELT-2 was the major TF controlling gut genes, such as *ges-1* (McGhee *et al.*, 2007). As the ELT-2 GATA TF has been identified as being essential for gut gene regulation in *C. elegans*, it is reasonable to presume that there may be a TF having a similar effect in *H. contortus*. Couthier *et al.*, (2004) identified an *elt-2* homologue in *H. contortus* that has similar expression both temporally and spatially to that observed in *C. elegans*. Moreover, ectopic expression of *Hc-elt-2* in *C. elegans* embryonic cells resulted in an increase in gut cells. The demonstration of functional

conservation of ELT-2 in gut development in these two species suggests that *Hc-elt-2* is also important for gut gene expression in later stages.

As well as experimental studies in *C. elegans*, parasite promoters can be analysed bioinformatically for motifs identified as being important for gene regulation in *C. elegans*. In addition, more general TF databases, such as TRANSFAC (<http://www.biobase-international.com/product/transcription-factor-binding-sites>), MEME (<http://meme.sdsc.edu/meme/cgi-bin/meme.cgi>) and LocalMotif (<http://en.bio-soft.net/dna/localmotif.html>) can search parasite promoters for any putative conserved motifs. However, caution must be exercised in such analysis as motifs may occur by chance and not be relevant to gene control. For example, putative binding sites for the DAF-16 TF (Gao *et al.*, 2010) have been identified in a large number of genes but the relevance of this is not known.

In the previous chapters, a number of multigene families were identified and annotated in *H. contortus*. Transcriptome and EST data has been used for analysis and to confirm expression of these genes. It is of interest to determine if any patterns or motifs are present in the promoter regions of these genes and progress our understanding of parasite gene regulation.

The main aims of this chapter were to;

- Carry out bioinformatic promoter region analysis of previously identified *H. contortus* CBL genes.
- Generate transgenic *C. elegans* worm strains containing *H. contortus* gene promoters linked to GFP and/or Lac-Z and use these in TF RNAi studies.
- Generate an integrated *C. elegans* transgenic strain containing a gut promoter fused to GFP as a tool to study gene control.



## 5.2 Results

### 5.2.1 Bioinformatic analysis of *H. contortus* promoters

#### 5.2.1.1 *H. contortus* gut gene promoter motifs

As the 5' upstream promoter region of genes has been identified as containing binding sites for TFs, these regions were analysed for any potential TF binding motifs. The upstream promoter regions for a number of the *H. contortus* CBL gut-expressed genes, discussed in previous chapters, were identified by annotation of the genomic data and in some cases from 5' RACE data. As the *H. contortus* genome has not yet been fully sequenced and assembled, upstream promoter region could not be identified for some genes and these were therefore omitted from this analysis. A number of motifs have previously been identified as potential regulatory motifs within promoters of *C. elegans* gut genes (Hawkins and McGhee, 1995). These include the GATA, TATA, CAAT and CANNTG (E box) motifs. Figure 5.1 is a diagrammatic representation of the upstream promoter regions of the *H. contortus* protease genes on BAC 18f22 indicating the location of these motifs. Studies in *C. elegans* have supported an important role for motifs close to the start codon. Work by MacMorris *et al.*, (1992) identified that just 200 bp of promoter region is required for expression of the *C. elegans* vitellogenin (*vit-2*) gene, for this reason analysis of the protease genes focussed on the proximal 1 kb upstream sequence.

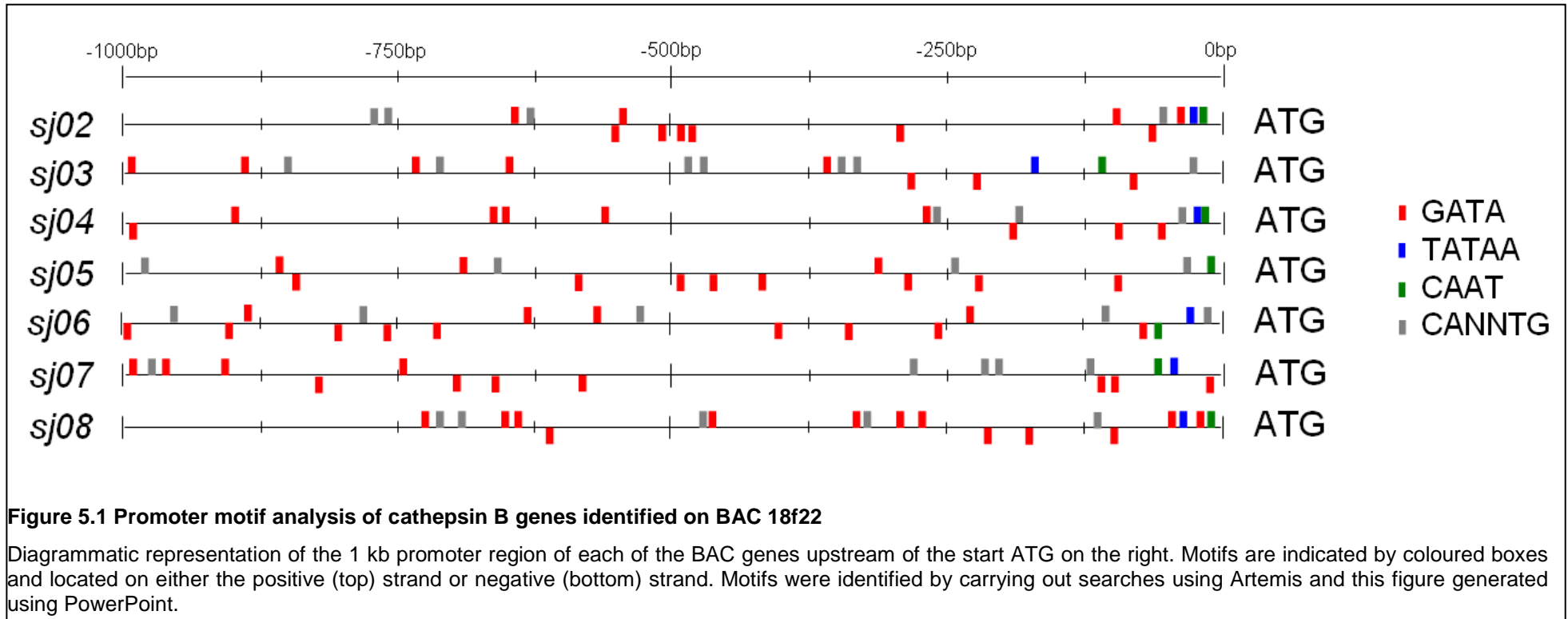
As previously mentioned, in *C. elegans* ELT-2 has been shown to be the major GATA-type TF regulating gut gene expression, therefore it was of interest to identify the number and positioning of GATA motifs within the *H. contortus* gut gene promoters. Within 1 kb upstream of the ATG start codon of the *H. contortus* AC protease family on BAC 18f22, all have a GATA motif on the negative strand at approximately position -85 bp. Additional GATA motifs either on the positive or negative strand were found further upstream in all of the AC gene promoters. Abundance of GATA motifs did not appear to correlate with level of expression, based on EST data (Table 5.1). However, further analysis of the BAC genes identified that for those genes in which no EST data is available (*sj02*, *sj07*, *sj08*), a GATAG motif is present, either on the positive or negative strand, between the start ATG and the GATAA motif at position -85 bp. Although

both GATAA and GATAG are consensus sequences, in this instance this GATAG motif could potentially be having a negative effect on gene expression. Work has been carried out looking at the effect of mutation of promoter GATA motifs and one study by Britton *et al.*, (1998) identified that the mutation of GATA motifs that are situated in the forward orientation resulted in a decrease in expression however mutation of those in the reverse orientation did not. Therefore, it may be speculated that in addition to the orientation of GATA motifs, the make-up of the consensus sequences may also have an effect on expression.

Analysis of the AC family gene promoters also identified a number of E box (CANNTG) motifs within 1 kb of the start codon. The first E box motif was within 23-120 bp upstream with positions indicated on Figure 5.1. It is unknown whether the position of the first E box motif upstream of the start ATG, or if the number of E box motifs within the 1 kb promoter region may have an effect on gene expression. In a study carried out by Zhao *et al.*, (2007) examining *C. elegans arg-1* gene expression, a GT motif (inverted E box motif GTNNAC) was identified that is thought to be important for motif function as its mutation resulted in a loss of gene expression. This inverted E box motif was found within 20-30 bp of other E box motifs. For this reason the 30 bp immediately upstream and downstream of all the E box motifs of the *Hc-AC* genes were identified and aligned in an attempt to identify any significant motifs or alignments that may be important for control of gene expression. Only one inverted E box motif was identified, upstream of the second E box motif in the *sj02* promoter indicating that this is not a common motif in the *Hc-AC* related genes. Interestingly, the first upstream E box motif in the *sj04* and *sj05* promoters have two GATA motifs within 30 bp upstream, a finding which may account for the observed higher number of associated ESTs associated with *sj04* and the reasonable number associated with *sj05*. Therefore, interactions between GATA TFs at closely positioned sites may enhance gut gene expression, particularly close to the start site. In contrast, genes *sj02*, *sj03* and *sj07* have E boxes rather than GATA motifs located within 30 bp of each other, which may have a less stimulatory effect, based on EST data (Figure 5.2).

Also included in Figure 5.1 is the location of the TATAA motif, positioned closely upstream of the start ATG. The TATAA motif, where present, tends to be located 25-30 bp upstream of the transcription start site and is associated with an

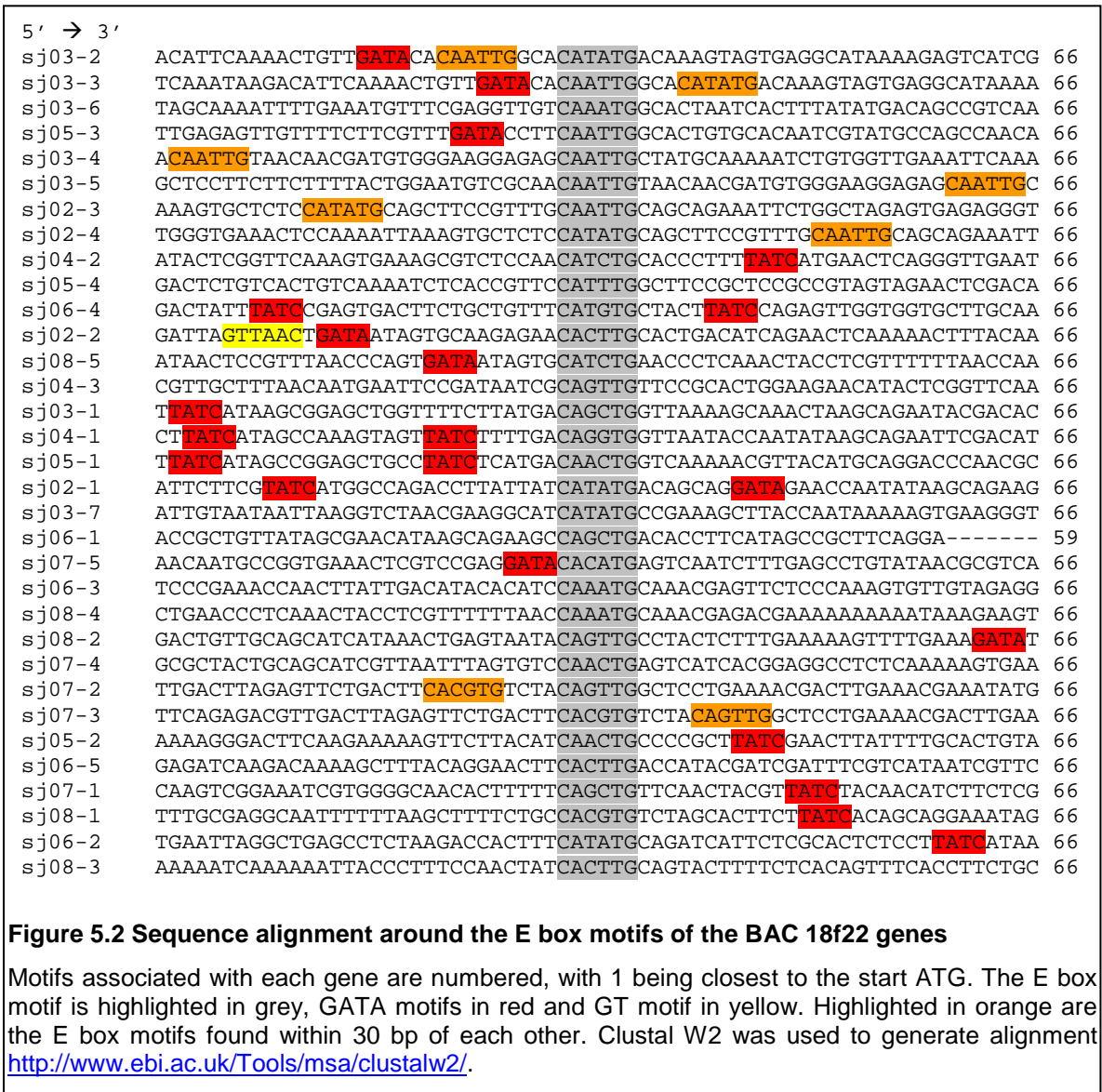
Initiator element (Inr) (Shi and Zhou, 2006). The position of the TATAA binding site is thought to be important as it determines the location of the transcription start site. As the function of the TATAA motif is also dependent on binding of transcription binding proteins (TBP) to surrounding sequences, the orientation of the TATAA motif sequence and subsequent location of these TBP can have an effect on transcription (Juo *et al.*, 1996). O'Shea-Greenfield and Smale (1992) showed that inversion of a TATAA motif led to less activity indicating that this motif is important for gene expression. Despite this evidence however, there are many promoters that do not contain TATAA elements and thus the TATAA motif may ultimately not be essential in certain gene promoters. One example *sj05*, has no TATAA motif close to the start ATG. However, EST data and RNA sequence data indicate expression of this gene, suggesting that the TATAA motif is not required, and for this gene the CAAT motif may be important for gene expression.



| Gene Family  | Number of GATA motifs | Position of TATAA motif | Position of CAAT motif | Number of CANNTG motifs | Nembase ESTs |
|--|-----------------------|-------------------------|------------------------|-------------------------|--------------|
| AC family on BAC 18f22   |                       |                         |                        |                         |              |
| sj02   | 10                    | -40                     | -37                    | 4                       | 0            |
| sj03   | 8                     | -191                    | -89                    | 7                       | 7            |
| sj04   | 9                     | -40                     | -28                    | 3                       | 72           |
| sj05   | 11                    | -                       | -11                    | 4                       | 7            |
| sj06   | 13                    | -48                     | -66                    | 5                       | 6            |
| sj07   | 11                    | -52                     | -67                    | 5                       | 0            |
| sj08   | 13                    | -51                     | -15                    | 5                       | 0            |
| HmCP family supercontig_0059492                                  |                       |                         |                        |                         |              |
| 1  | 9                     | -33                     | -121                   | 7                       | 15           |
| 2  | 8                     | -37                     | -124                   | 3                       | 0            |
| 4  | 9                     | -33                     | -50                    | 4                       | 23           |
| 5  | 4                     | -                       | -19                    | 2                       | 38           |
| 7  | 9                     | -33                     | -128                   | 5                       | 28           |
| gcp-7 like family on supercontig_0041161 and supercontig_0059702 |                       |                         |                        |                         |              |
| 2  | 6                     | -382                    | -93                    | 5                       | 6            |
| 5  | 14                    | -503                    | -99                    | 4                       | 0            |
| 6  | 7                     | -14                     | -72                    | 3                       | 6            |
| 7  | 5                     | -137                    | -113                   | 5                       | 0            |
| 8  | 2                     | -288                    | -136                   | 0                       | 10           |
| 10   | 11                    | -724                    | -44                    | 2                       | 10           |
| 11   | 10                    | -437                    | -45                    | 8                       | 3            |

**Table 5.1** Number of conserved motifs in the 1 kb promoter region of a number of *H. contortus* genes

Motif identification carried out using Artemis.

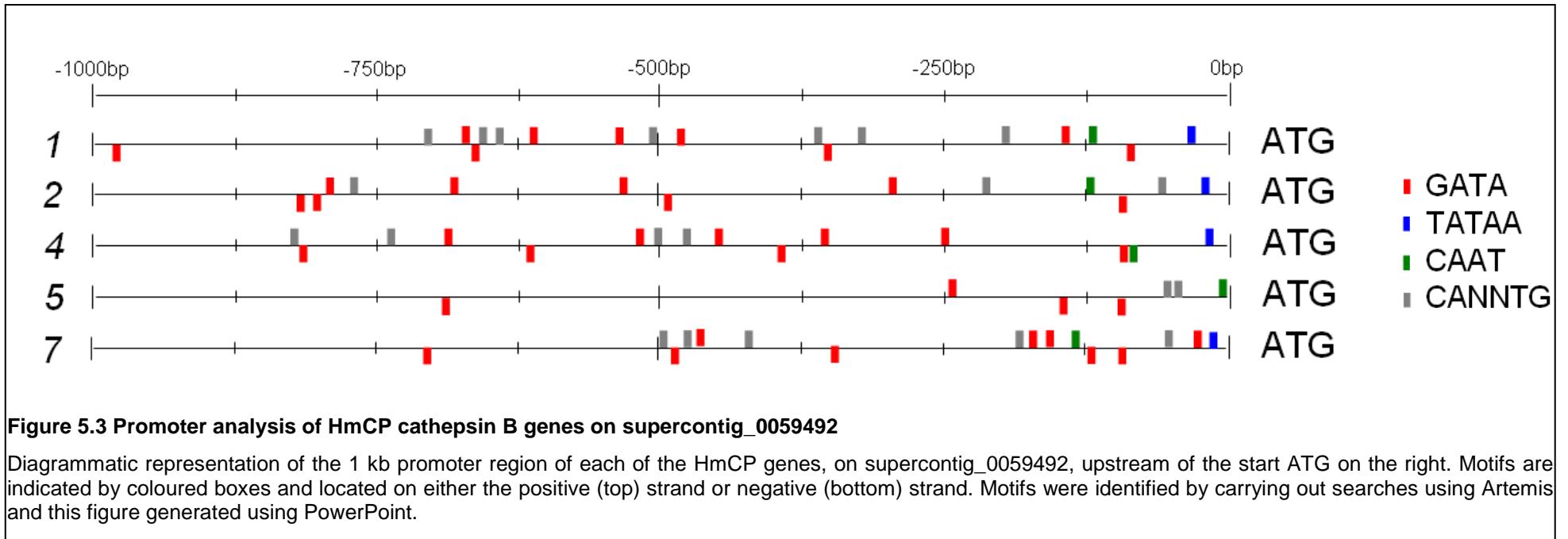


**Figure 5.2 Sequence alignment around the E box motifs of the BAC 18f22 genes**

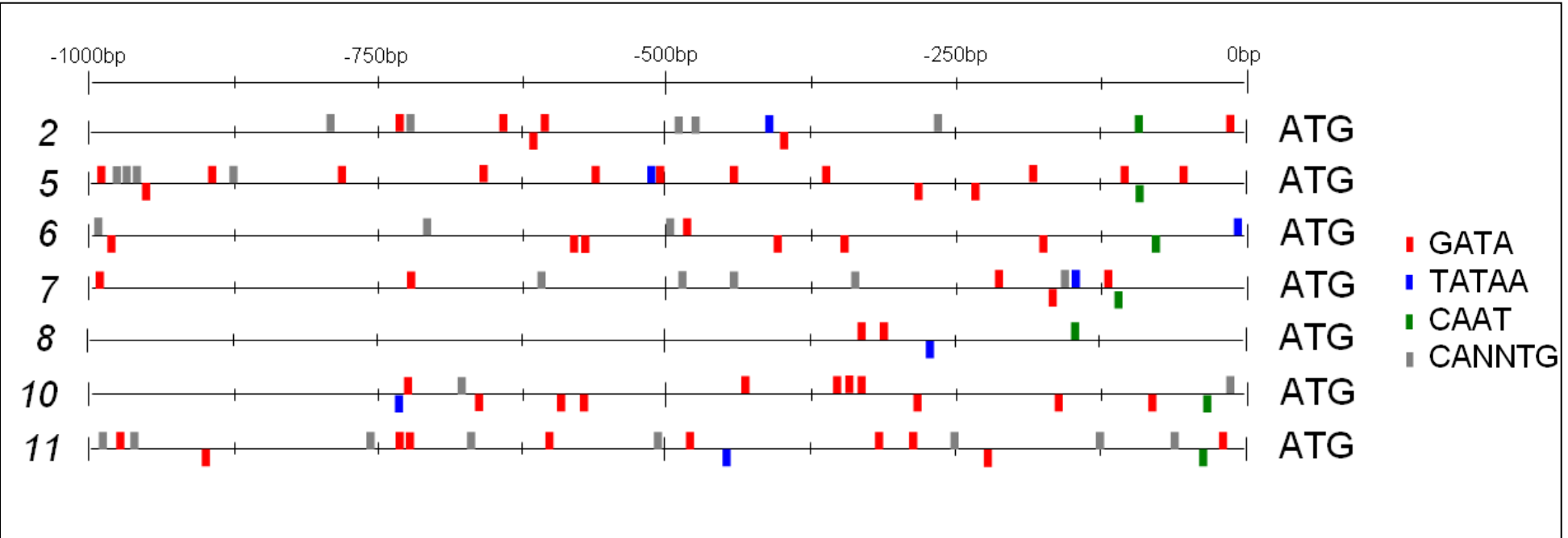
Motifs associated with each gene are numbered, with 1 being closest to the start ATG. The E box motif is highlighted in grey, GATA motifs in red and GT motif in yellow. Highlighted in orange are the E box motifs found within 30 bp of each other. Clustal W2 was used to generate alignment <http://www.ebi.ac.uk/Tools/msa/clustalw2/>.

Analysis of the *H. contortus* AC cathepsin B genes showed that they all have a TATAA and/or CAAT motif within 90 bp of the start codon. A TATAA or CAAT motif close to the start codon (19-37 bp upstream) was also identified in the promoters of HmCP cathepsin B genes on supercontig\_0059492 (Figure 5.3, Table 5.1). Of the five HmCP gene promoters analysed, the second is the only one with no associated ESTs, however there is no obvious explanation for this. Interestingly, most of the *gcp-7*-like genes on supercontig\_0041161 and supercontig\_0059702 contained no TATAA motif and where CAAT motifs were found, these were often further upstream of the start codon than in the AC genes (Figure 5.4, Table 5.1). The region surrounding the first E box motif upstream of the start ATG was analysed for both the HmCP family and *gcp-7*-like genes (Figure 5.5). Interestingly, for the HmCP genes, the fourth, fifth and seventh, which have a high number of associated ESTs there is either a GATA or

another E box motif found within 30 bp of each of these. This feature is not found in the first (15 ESTs) or the second (0 ESTs) of the HmCP genes. However, it is difficult to generalise that GATA motifs close to E box sequences enhance activity as this situation exists for some of the *gcp-7*-like genes, which have low or no associated ESTs.







**Figure 5.4 Promoter analysis of *gcp-7*-like genes on supercontig\_0041161 and supercontig\_0059702**

Diagrammatic representation of the 1 kb promoter region of the *gcp-7*-like genes upstream of the start ATG on the right. Motifs are indicated by coloured boxes and located on either the positive (top) strand or negative (bottom) strand. Motifs were identified by carrying out searches using Artemis and this figure generated using PowerPoint.

## HmCP family on supercontig\_0059492

```

5' → 3'
2  T TATC GGATTTAAGACTAATCCTTATATGACAGCTGACTTAATGAATATAAAATAAGGAGATGCTCT 66
7  TCGGAAGTCAAAGTCTAATCCTTAAATGACAGCTGAGCGTAACGAATATAAGGA GATA CTCACGT 66
5  AGGCTAATCCTCCCATGCAGT CACCTG ACCCAAATGAATTTAAGGAGGTGCTGACTTGGCCTAATT 66
1  GATGACGGGTTCAACTCAAAAAACCATGACC AAGTGAAATCCTTAAGTAAACTGTGAAACGACTTT 66
4  GATA GTGCGC CATTTC CCACTACACCATCTCAGGTGAACCGTTTCCCGCGCGTTCGTCGCGCGACC 66

```

## gcp-7-like on supercontig\_0041161 and supercontig\_0059702

```

5' → 3'
7  ACTGATTAT TATC GAATGTTTTTGTCTATGCCACATGCCTTTTTATAAAGTCATCTAGA GATA AGGC 66
11 CTTCACTTATTCTTTTTGCGACTATTTATTCATTTGCTACTGATGTAGAAATGAGGC GATA ACAT 66
6  TGAAATGGAATGTATATGGTTTTTCTTTTTCCATATGACATGACTCCT GATA AAGTTCCTCTGTGACA 66
10 CCCC GTTAATGGCGATTGCTTTACTTATTCATTTGAGGAAAAAGTTGACATAATAAG 58
5  GGCAC TC TATC CCTGTACGAAGGCCAAAGCCCAAATGTCATGC GATA GAATTGTTTGTAAATTTGTTG 66
2  TAAGTCTTTTCAAATAAGAGTAGTCAATGCATATGTGGCCCATGAAATGACTTGGTGGGGCCGGT 66

```

**Figure 5.5 Sequence alignment around E box motifs of the HmCP family and gcp-7-like genes**

The E box motif is highlighted in grey, GATA motifs in red and the E box motifs found within 30 bp of each other highlighted in orange. Clustal W2 was used to generate alignment <http://www.ebi.ac.uk/Tools/msa/clustalw2/>.

The *cpr-6* gene promoters for both *H. contortus* and *C. elegans* were also analysed for conserved motifs. CPR-6 proteases have a high level of conservation but despite this there is little sequence conservation within the promoter regions. Alignment of both 1 kb and 2 kb promoter regions was carried out and interestingly, when 2 kb of upstream region for each species was aligned there was a much higher percentage identity (Figure 5.6). Spaces were left in the alignment to ensure the best fit, which suggests that in *H. contortus* the promoter region may have extra sequence with binding sites potentially being more dispersed. This is consistent with the larger genomic size of *H. contortus* and increased intron lengths compared to *C. elegans* (Laing et al., 2011). While the previously mentioned promoter motifs, GATA, TATA and CANNTG could all be identified, the location of these was not conserved between *C. elegans* and *H. contortus*. As found with some of the other *H. contortus* cysteine protease genes analysed, these motifs are located reasonably close to the start ATG of *Hc-cpr-6*, with the first GATA, TATA, CAAT and CANNTG motifs being identified between positions -12 to -65 bp. However, in *C. elegans* these motifs are located further from the start ATG, between positions -64 to -335 bp. This was unexpected as even with the addition of sequence to the *H. contortus* promoter,

it appears that these motifs are located closer to the start ATG than in *C. elegans*.

|                     |      |  |
|---------------------|------|--|
| <i>C. elegans</i>   | -340 | CTG-----ACATCTGACTTTTT-----ATTAGGTTTTTCCAT-CA                  |
| <i>H. contortus</i> | -624 | TCCGAGGAGGTAAGCCCTTAGCTTCTTACTCTTTGAGAAAAATAAAGGGAAACCATGCA    |
|                     |      | *                        |
| <i>C. elegans</i>   | -306 | TAT--AACCCTTTCAAACGAAATTA-----ATGTGCTAAAT                      |
| <i>H. contortus</i> | -564 | AATGGAAAACAAGTGACGCACTTGAGAGAATCCTAAGGTCGGATCCACGATGATCTTAAA   |
|                     |      | *                        |
| <i>C. elegans</i>   | -272 | CTG---TTAAGTTTTCAATATT-----TTCCTTGTCTTTAGGT-----CAATCT         |
| <i>H. contortus</i> | -504 | AAAGACTTAGATTTCAATCATTTCAGAAAATACAAGAACTTTTTGTGTACTTTTCTTTCC   |
|                     |      | *                        |
| <i>C. elegans</i>   | -232 | TCTTTGCCACACAG-TTCAAACACTACT-----ACCGCCGAGTCACGT-----          |
| <i>H. contortus</i> | -444 | TCTTAGCTATTTTCGCTTCAATAGCTTTGGTGATCTCCTGCAGGTCTTGAGAATTTGTAA   |
|                     |      | *                        |
| <i>C. elegans</i>   | -192 | ----CACACCATCACAGGATAGTGACCGGTCTAG-----GATGTAC-----CCTGACAC    |
| <i>H. contortus</i> | -384 | TATGTACTACTCCCTTTGGACAGCTCGACGTTCTAGCTATTGGTGAACGAACCTACAATTCA |
|                     |      | *                        |
| <i>C. elegans</i>   | -146 | TGTGATGGACGCAGCCG-----ACACTCTTATCGAAATGCACAGGGCC-----          |
| <i>H. contortus</i> | -324 | TGGAATAGGAGAATCCAGAAAAGCGATGCTAAAACATCCATATTCAAGGGGCCGAAGTGC   |
|                     |      | *                        |
| <i>C. elegans</i>   | -103 | -----AAATTTGATAACG---AAAACATGTTCTATAAAAG-----                  |
| <i>H. contortus</i> | -264 | TAACGTAAGATCTCTGTGGAAATTTGTTTTTCGGCGGTGATCTGATATGGGAACGGGTCAC  |
|                     |      | *                        |
| <i>C. elegans</i>   | -71  | -----CATG---CTGATAAAAGCG-----AGCAGTCAAGC-----GACGAC---         |
| <i>H. contortus</i> | -204 | TTCTTCATGCTCTTATCAAACCGTACCTGACCGTCAATCTCCCTGATAAAGTCGTCTCG    |
|                     |      | *                        |
| <i>C. elegans</i>   | -38  | ----AACTTGCGATCAACACGCTGACCGTCG--ACGCCAACATG                   |
| <i>H. contortus</i> | -44  | TATAAAATAGCCGACAACAAGTGGTCTTCAGTTGGACACGATG                    |
|                     |      | *                        |

**Figure 5.6 DNA alignment of the *H. contortus* and *C. elegans cpr-6* promoters**

A segment of the 2 kb promoter alignment closest to the start ATG containing the promoter motifs. GATA motifs are in red, TATAA in blue, CAAT in green and E box motif in grey. Conserved bases are indicated by a star. Clustal W2 was used to generate the alignment <http://www.ebi.ac.uk/Tools/msa/clustalw2/>.

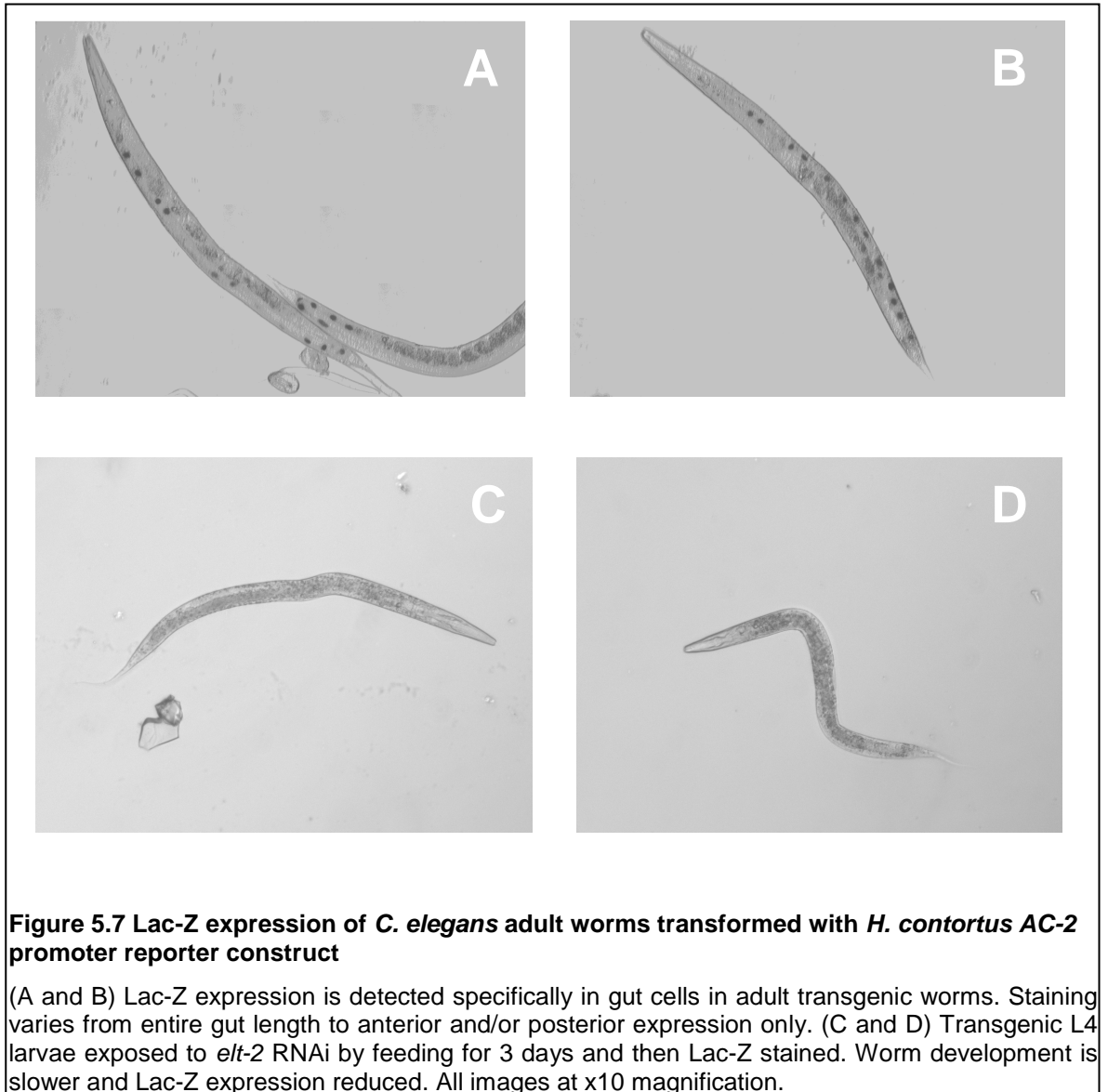
In addition to the conservation of regulatory motifs within promoter regions, it has been suggested that the presence of repeat sequences either within the promoter region, upstream to this or within intronic regions may have an effect on gene expression (Liu *et al.*, 2002). Work carried out in Chapter 3 identified transcript data for all the genes present on BAC 18f22 except for *sj08*, suggesting that *sj08* is not expressed. Analysis of the promoter regions of the AC genes on the BAC identified a very large repeat sequence 1.5 kb upstream of the *sj08* predicted start ATG. The repeat sequence is approximately 1,200 bp in length and has a conserved ‘ACAGACAGACAG’ pattern. As yet there is no other repeat sequence identical to this identified in the *H. contortus* genome. Work carried out by Liu *et al.*, (2002) on *Mycoplasma gallisepticum* showed that the presence of a trinucleotide (GAA)<sub>12</sub> repeat, in the promoter, was essential for expression of the pMGA gene. As well as repeat sequences being essential, it is also possible that repeat sequences at specific locations may have a negative effect on gene expression.

In this study a number of motifs potentially involved in TF binding have been identified in the *H. contortus* protease genes, some close to the start and occasionally clustered. Further genome data as well as comparative genomics, will aid in determining if these are likely to be important for gene expression.

## 5.2.2 Practical analysis of gut gene expression

### 5.2.2.1 Expression pattern of *H. contortus* *sj04* (AC-2) promoter in transgenic *C. elegans*

EST analysis using Nembase (<http://www.nematodes.org/nembase4/blast.shtml>) identified 72 ESTs for *Hc-AC-2*, higher than any of the other BAC genes. For this reason it was chosen to experimentally test expression in *C. elegans*. Previous work has already been carried out by Britton *et al.*, (1999) examining expression of this (AC-2) gene in *C. elegans* and looking at the effect of removing a GATA motif in the promoter region. In the previous study 2.3 kb of upstream sequence was inserted into the Fire vector pPD95.03 which contains a *lac-Z* reporter (Fire *et al.*, 1990). In an attempt to increase the level of expression and enable analysis in live worms the newer Fire vector pPD96.04 was used as this contains both GFP and Lac-Z reporters. In addition to this the first intron of *Hc-AC-2* was included in the construct as it has previously been demonstrated by Ho *et al.*, (2001) and Fire *et al.*, (1990) that intron inclusion can increase gene activity in *C. elegans*. Primers were designed to amplify a 1 kb promoter region and intron one (primers in Appendix 2 Table 5.1). This construct in pPD96.04 was injected in conjunction with a *rol-6* marker gene, into the gonad of *C. elegans* wild-type worms (25 ng/μl and 100 ng/μl respectively). GFP expression was not demonstrated for any of the transgenic lines, but they were Lac-Z positive after staining of larvae and adult stages. The majority of the worms displayed staining restricted to the anterior and posterior regions of the gut however in some worms staining was observed down the entire length of the gut (Figure 5.7 panel A and B).



### 5.2.1.2 RNAi of *Ce-elt-2* in *C. elegans Hc-AC-2* transgenic worm strain

As mentioned previously, the *AC-2* promoter region contains seven GATAA motifs in the 1 kb promoter region and may therefore be positively regulated by a GATA TF. To determine whether the *C. elegans* ELT-2 GATA TF can control *H. contortus* gut gene expression, *elt-2* RNAi of the *Hc-AC-2::lac-Z* transgenic strain was carried out using the RNAi feeding method (Chapter 2.1.5). Any decrease in Lac-Z expression was then examined visually. RNAi of *elt-2* in the *Hc-AC-2* promoter strain shows reduced expression, in most worms staining of Lac-Z was absent or only in 1-2 gut cells (Figure 5.7 panel C and D). This is in contrast to worms fed on L4440 control plasmid where approximately 85% showed expression throughout the gut. Attempts were made to quantify Lac-Z expression by RT-PCR but this proved difficult.

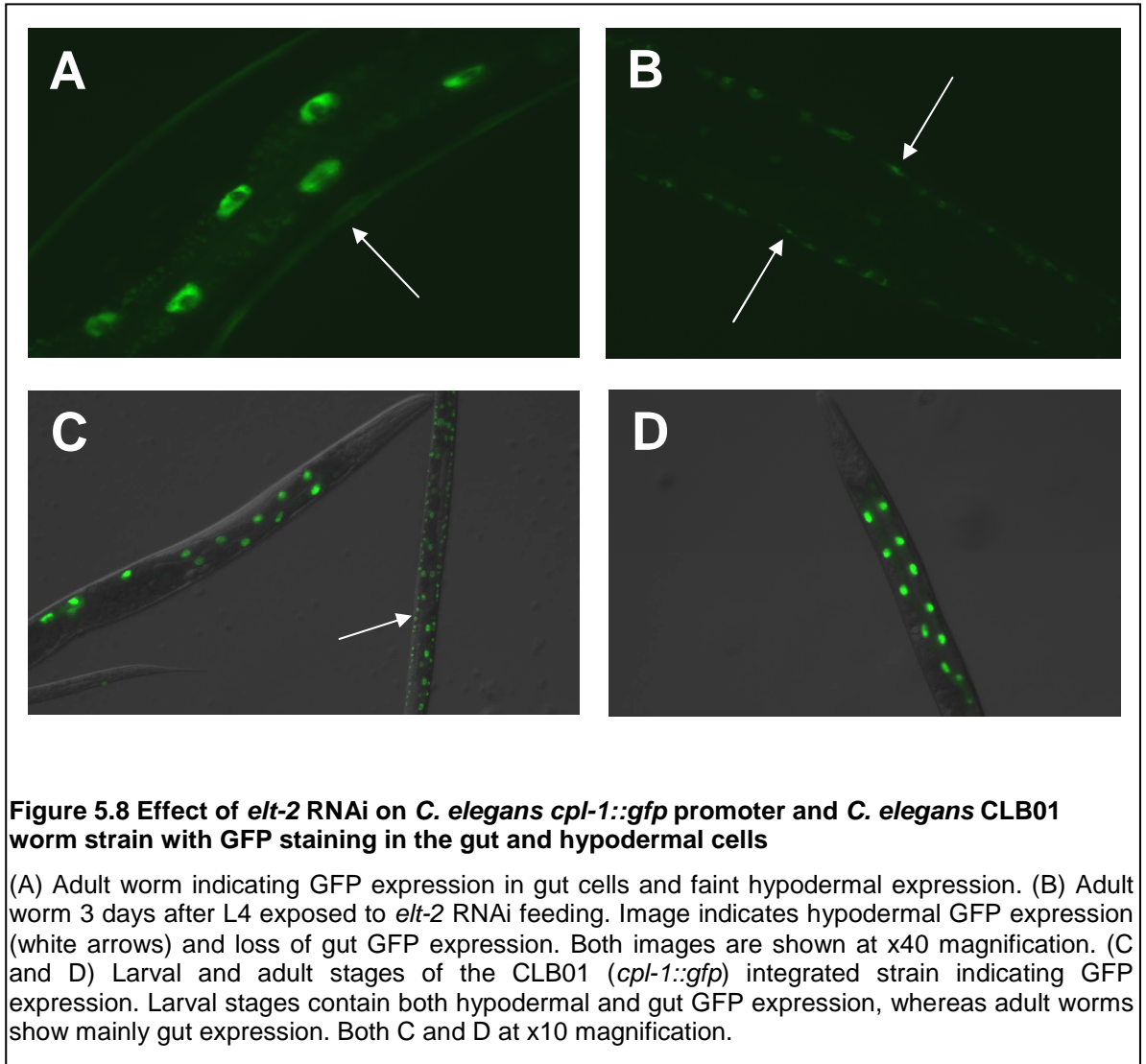
### 5.2.1.3 RNAi of *Ce-elt-2* in *C. elegans Ce-cpl-1::gfp* transgenic worm strain

The *C. elegans* transgenic worm strain containing the *H. contortus AC-2* promoter exhibited mosaicism between individual worms. In addition, no GFP expression was detected and therefore analysis of living worms was not possible. Significant work has been carried out examining *cpl-1* gene expression in both *C. elegans* and *H. contortus* (Britton and Murray, 2002). This is a cathepsin L-like cysteine protease gene that is required for embryonic development and accumulates in the adult *C. elegans* gut with some additional expression in the hypodermal cells. Hashmi *et al.*, (2002) generated a *Ce-cpl-1* reporter construct containing 1.7 kb of promoter sequence in Fire vector pPD95.75 (Fire *et al.*, 1990), that was co-injected with an *unc-76* rescue plasmid to generate a transgenic *C. elegans* worm line. This *C. elegans cpl-1* promoter::*gfp* strain was used to determine whether *Ce-elt-2* RNAi could cause a clear reduction in GFP expression. *elt-2* RNAi feeding of L3/L4 stage transgenic larvae was carried out as described for the *H. contortus AC-2* reporter strain (Section 5.2.1.2). This resulted in significant reduction or complete loss of *gfp* in gut cells, in contrast to worms fed control plasmid (Figure 5.8 panels A and B). This confirmed that ELT-2 GATA TF is required for *cpl-1* expression in *C. elegans* gut cells. GFP expression could still be observed in hypodermal cells, particularly in larval stages, suggesting involvement of another TF, possibly GATA-3 (Spieth *et al.*,

1991). This also confirmed a specific effect of *elt-2* RNAi in gut cells and not just a general reduction in gene expression.

#### **5.2.1.4 Generating an integrated *C. elegans cpl-1::gfp* strain as a drug screening tool**

The specific and significant effect of *Ce-elt-2* RNAi on the *C. elegans cpl-1::gfp* strain as well as the ease of working with live GFP expressing worms indicated that this strain would be useful for screening compounds as potential inhibitors of ELT-2 function and testing for a decrease in GFP level (Chapter 6). This transgenic strain has a transmission rate of approximately 70% and therefore to make expression studies easier an integrated line was developed (Chapter 2.1.4). After integration, the transgene was present in 100% of the worms and strong GFP fluorescence was visible in the gut and hypodermal cells (Figure 5.8 panels C and D). To ensure that the *Ce-cpl-1* insert was still present in the integrated strain (named CLB01), single worm lysis PCR was carried out on both the integrated strain and on wild-type worms as a control. Gene specific and vector specific primers were used (Primers in Appendix 2 Table 5.2). All the integrated worms produced specific bands with both sets of primers, while wild-type worms only produced a band with the gene specific primers, indicating that all integrated worms carried the transgene. This integrated *C. elegans cpl-1::gfp* strain could then be used in drug screening assays (Chapter 6).





### 5.3 Discussion

The aim of this chapter was to examine the regulation of gut gene expression in *H. contortus* using both bioinformatic and practical analysis. Prior to this, only limited analysis of selected *H. contortus* promoter regions had been carried out. However, with the availability of genome data combined with gene annotation, it is now possible to examine and compare promoter sequences for any conserved patterns.

Hawkins and McGhee (1995) identified that GATA TFs are important for the regulation of gut genes in *C. elegans* and more recent work by Couthier *et al.*, (2004) suggested that the same may be true in *H. contortus*. Bioinformatic analysis of the *H. contortus* gut gene promoters carried out here identified an abundance of GATA motifs in the 1 kb promoter region, with some genes having up to seven GATA motifs. Interestingly however, there does not appear to be any correlation between number of GATA motifs and gene expression. For example, *sj04* with seven motifs has 72 associated ESTs and *sj06* also with seven motifs has only 6 associated ESTs. This finding is perhaps not surprising as examination of the promoter regions of *C. elegans* genes, *Ce-pho-1* (Fukushige *et al.*, 2005) and *Ce-cpr-1* (Britton *et al.*, 1998), identified that where a number of GATA motifs were present in the upstream region, deletion studies identified that only one of these motifs was essential for gene expression. The current work raised the possibility that positioning of GATA motifs (and possibly E box motifs) relative to one another may influence level of expression.

Numerous studies examining the expression of parasite genes and gene promoters in *C. elegans* have concluded that they are functional and useful for comparative genomics (Britton *et al.*, 1999; Gomez-Escobar *et al.*, 2002; Kwa *et al.*, 1995). While *C. elegans* promoter-reporter constructs most often produce robust expression, consistent in the majority of transgenic worms, parasite promoters expressed in *C. elegans* have been found to be expressed at lower levels. This most likely reflects lack of complete conservation of transcription factors and motifs across nematode species and while heterologous expression is possible, it may not be optimal. In this study, Lac-Z staining was required to detect any expression from the *Hc-AC-2* promoter construct, which is more

sensitive than GFP visualisation. However, Lac-Z staining results in worm losses throughout the process and also requires fixation and therefore death of the worms. In contrast, the *C. elegans* CLB01 *cpl-1* promoter::*gfp* integrated strain generated in this work expresses GFP strongly in all worms, making subsequent RNAi or drug treatment assays easier and more reliable.

Studies aimed at using *C. elegans elt-2* RNAi to examine whether *H. contortus* AC-2 is regulated by the ELT-2 GATA TF produced variable results which were difficult to reliably quantify. An alternative approach was also taken to examine *Hc-AC-2* regulation. This made use of a construct termed pAC9 (Couthier *et al.*, 2004), consisting of a *Hc-elt-2* cDNA clone inserted downstream of hsp16-2 heat shock promoter, in vector pPD49.78 (Fire *et al.*, 1990). This worm strain has the ability to over-express *Hc-elt-2* after heat shock at 33.5°C for 2 hours. The pAC9 construct was injected into the *Hc-AC-2* promoter transgenic *C. elegans* worms (Section 5.2.2.1) together with *myo-2::gfp* pharyngeal marker. Positive lines were obtained and were exposed to the heat shock process to determine if any increase in *Hc-AC-2* expression could be observed, which would demonstrate that the *Hc-AC-2* promoter is regulated by the ELT-2 GATA TF. Unfortunately no increase in gene expression occurred. This finding could be due to there already being a high level of *elt-2* in the transgenic worms and therefore even with increased *elt-2*, no enhanced expression occurs.

RNAi of *C. elegans elt-2* was carried out using both the feeding and soaking method. Of the two approaches, feeding produced the most reliable reduction of *C. elegans cpl-1::gfp* expression. This could be due to worms not being as healthy while exposed to dsRNA in liquid and also to the shorter exposure time which could be applied due to worms starving after an extended period. In *C. elegans* ELT-2 is required for gut development (Fukushige *et al.*, 1998). As previously indicated, the ELT-2 GATA TF has been identified as important for the study of gene expression in both *C. elegans* and *H. contortus* (Geldhof *et al.*, 2006). *elt-2* is required for gut development and thus removal of this results in worms that are smaller and starved. As noted by the RNAi results, *elt-2* RNAi efficacy varies depending on the stage of the worm. It was identified that initial exposure of worms at the L1-L3 stages, show the greatest *elt-2* RNAi effect. This is in comparison to testing L4 and adult stages in which the RNAi effect was less severe. The reasoning for this may be due to the requirement of *elt-2* for gut

differentiation (Fukushige *et al.*, 1998) and it is also the main TF regulating expression of functional gut genes during larval development (McGhee *et al.*, 2007). In the work carried out here, more reliable reduction of *Ce-cpl-1::gfp* could be monitored by using transgenic worms at the L3/L4 stage in RNAi feeding assays. In these worms gut formation as well as expression of gut genes have already occurred and healthy worms develop. In contrast, when embryos or L1-L3 stage larvae were applied to *elt-2* RNAi plates, abnormal, smaller and starved worms were observed and monitoring of *cpl-1::gfp* expression was difficult. Therefore interfering with *elt-2* activity in early development has a significant detrimental effect on gut function. If the same is true in parasitic species, targeting ELT-2 is likely to have severe consequences.

Another putative role of *elt-2* is in the *C. elegans* immune response to bacteria. Reece-Hoyes *et al.*, (2005) identified 934 TF genes in the *C. elegans* genome and applied strict criteria to this list to identify those thought to be important in immunity. ELT-2 was one of the five identified and experimental work indicated a decrease in immunity after *elt-2* RNAi. As *elt-2* is required for normal intestinal differentiation it was important to determine that the observed results were not due to the intestinal effect of the *elt-2* RNAi. An immune effect was confirmed and therefore *elt-2* ablation is a realistic target for the control of nematode worms, both through an increased susceptibility to bacteria and due to a decrease in intestinal growth and development.

In conclusion, bioinformatic analysis of parasite gut gene promoters has identified that GATA motifs are over-represented in the 1 kb upstream region. This feature may be significant as previous work by Britton *et al.*, (1999) has identified that gene expression can be abolished with the removal of these motifs. Additionally, it was demonstrated that expression of gut gene promoter-reporter constructs can be reduced by *Ce-elt-2* RNAi. The ability to interfere with gut development and gut gene expression is useful for future parasite control as compounds that have the same effect as *elt-2* RNAi may provide an alternative to current parasite control methods.

## **Chapter 6**

Screening for potential inhibitors of gut gene  
expression in *C. elegans*

## 6.1 Introduction

The increasing problem of anthelmintic resistance has led to the requirement for new effective anthelmintic drugs. Traditional anthelmintic drug discovery methods have relied on whole parasite screening. This method of screening involves testing a number of compounds against whole parasites to identify active compounds and subsequently identifying the target (Geary *et al.*, 2004). For parasitic nematodes, whole organism based screening has the advantage of only focussing on compounds taken up by the parasite. The three main classes of current anthelmintic drugs, as well as the amino-acetonitrile derivatives (AADs), were discovered using this approach and their mechanism of action subsequently identified using *C. elegans*. The use of model organisms in chemical screening relies on the presumption that there will be high biological conservation between the model organism and the parasite of interest. *C. elegans* has been used as a model to identify not only anthelmintic targets but also resistance mechanisms in parasitic species, focusing on molecules conserved between this organism and the parasite (Gilleard *et al.*, 2005). For example, the macrocyclic lactones were shown to affect glutamate-gated chloride channels, causing paralysis of the nematode due to an increased permeability of the muscle to Cl<sup>-</sup>. Work using *C. elegans* enabled the protein structure to be used to identify the inter-species conservation of motifs in the ligand-gated ion channels (Cully *et al.*, 1996). Another advantage of using *C. elegans* is the ease of generating mutants to allow identification of the mode of action for drugs. For example AAD-resistant *C. elegans* mutants were generated using ethane methyl sulphonate (EMS) mutagenesis (Kaminsky *et al.*, 2008b). Of the 44 AAD-resistant mutants studied, 36 had mutations within the *acr-17* and *acr-23* genes. These genes belong to the DEG-3 nAChR subgroup and this work resulted in identification of the possible mechanism of action for the AADs (Kaminsky *et al.*, 2008a).

More recently, for human drug discovery, cell screening methods have been replaced with mechanism based approaches in an attempt to identify compounds interacting with specific targets. The shift to mechanism based screening allows chemists to use structural information on the target to identify novel drugs, where the target is already identified (Woods and Williams, 2007). However this

requires a significant amount of work to identify appropriate targets and purify these in sufficient amounts for high throughput screen.

Previously, screening compounds for human diseases has also relied on non-targeted methods. For example Tamoxifen, the commonly used breast cancer drug which binds to oestrogen receptors, was discovered in the 1960s by the traditional screening method of testing a number of compounds against the disease of interest. It was first identified in a screen designed for new contraceptive drug discovery, however further insight into its mechanism of action identified it as being an effective anti-cancer drug (Jordan, 2003). Many drugs were initially identified in this way, because of their biological effects in certain screens. In many cases it was not until after their discovery that the mechanism of action and specific effects was identified (Latchman, 2000). Targeted screening has the potential advantage of designing drugs to specifically act on disease-associated molecules.

There are a number of key requirements when developing new anthelmintic drugs; they must have broad-spectrum activity, be inexpensive to manufacture, safe and easy to dose, and for livestock, have short withdrawal periods (Lanusse and Prichard, 1993). Additionally, anthelmintic drugs must be toxic solely to the parasite and not interfere with host function. The identification of new anthelmintic targets will be dependent upon the parasite having a higher susceptibility to the drug, or drug being more concentrated in the parasite target area than the host (Geary *et al.*, 1999).

Large scale screening is routinely carried out in industrial environments. The majority of the experimental work for this chapter was carried out at Pfizer in Kalamazoo, Michigan. Small molecule compounds available for screening at Pfizer are located in the RGate database. This is an internal company database that contains approximately 3 million compounds. It provides information about the structure of each compound, any screens on humans or parasites that have previously been carried out and the results of experimental work. Screening is regularly carried out on a number of parasites including *H. contortus* L3 larvae. This parasite is widely used for anthelmintic screening as it is available relatively cheaply, in bulk and L3 stage *H. contortus* can be stored for several months. At

Pfizer, screening with *C. elegans* is less favoured as the data does not always correlate well with activity against parasitic nematodes.

Transcription factors have been identified as potential drug targets, for example of anti-asthma drugs. Asthma is a chronic disease, characterized by increased expression of cytokines which in turn causes the increased activation of downstream TFs. The resulting effect of this TF stimulation is the increased expression of inflammatory genes (Adcock and Caramori, 2001). Corticosteroids are commonly used in anti-asthma therapy and act through TF inhibition (Barnes and Adcock, 1998). An aim of the current work was to determine whether it is possible to interfere with gut gene expression in nematodes in this way. Previous work identified that GATA TFs are important in the regulation of gut gene expression in *C. elegans* (Hawkins and McGhee, 1995).

In this project, families of *H. contortus* protease genes potentially regulated by GATA motifs and expressed in the gut have been identified. A number of studies in *C. elegans* led to the identification of the ELT-2 GATA TF. Subsequent work has shown that *elt-2* displays gut-specific expression with the ELT-2 protein being present in the nuclei of all gut cells. Additionally, it was shown that loss of *elt-2* function is lethal in embryos and early larvae due to intestinal malformation (Fukushige *et al.*, 1998). For this reason, targeting the ELT-2 GATA TF could potentially affect the expression of gut genes affecting gut development and function. The integrated transgenic *C. elegans* strain (CLB01) expresses GFP under the control of the *C. elegans cpl-1* promoter and directs fluorescence in the gut and hypodermal cells. RNAi of *Ce-elt-2* (Chapter 5) suggested that this promoter is controlled by the ELT-2 GATA TF and thus GFP fluorescence from this reporter construct can be used as a marker of ELT-2 TF activity.

The main aims of this chapter were to;

- Generate an *in vitro* assay to screen compounds using the integrated CLB01 transgenic *C. elegans* worm strain.
- Identify compounds that interfere with gut gene expression and potentially target the ELT-2 GATA transcription factor in *C. elegans* worms.
- Screen compounds in *H. contortus* and look for effects on parasite survival and motility, as well as characterise the specific effects of identified compounds.



## 6.2 Results

### 6.2.1 Determining the conditions required for screening *C. elegans* worms

Initial drug screening was carried out at Pfizer, Kalamazoo, Michigan, using *C. elegans* worms in liquid medium in multi-well plates. As *C. elegans* is not routinely used by Pfizer for screening, all parameters had to be optimised. These included, sample volume, volume of OP50 bacteria food source, number of worms and concentration of DMSO, as each can influence the measurements recorded and worm viability. The initial sections of this chapter describe work required to establish a suitable assay.

The CLB01 *C. elegans* worms show green fluorescence in gut and hypodermal cells. Prior to this work the level of fluorescence had not been measured. The EnVision 2103 Multilabel Reader was used to measure fluorescence levels. A blacked out 96 well plate was used for sampling to block out external light. Fluorescence readings were measured from the top of the wells as there was no plastic surface to penetrate. The settings used for the EnVision programme were; excitation filter FITC 485, emission filter FITC 535, a measurement height of 2 mm, excitation light of 5% and a detector gain of 10%.

The first fluorescence reading of CLB01 worms in 200  $\mu$ l of M9 buffer was recorded. This however was an excessive final volume and resulted in the worms being more dispersed, within individual wells, resulting in inaccurate readings. 100  $\mu$ l of CLB01 worms that were mixed stage and healthy, in M9 buffer, were then tested. As this was a more workable sample volume the readings should be more consistent and accurate. This sample condition was compared to: 100  $\mu$ l of N2 worms that were mixed stage and healthy in M9 buffer, 100  $\mu$ l of OP50 alone and 100  $\mu$ l of M9 buffer alone. From these first samples the wells containing CLB01 worms gave higher readings than two of the other conditions. Unexpectedly high readings were observed for the OP50 wells with the fluorescence reading much higher than that for the wells containing CLB01 worms.

To determine if the initial results observed were reliable, a 50  $\mu\text{l}$  sample from a 100  $\mu\text{l}$  well containing the CLB01 worms was transferred to a different well. M9 buffer was added to a total volume of 100  $\mu\text{l}$  in each well and 5  $\mu\text{l}$  of OP50 added only to the second well. The fluorescence reading in the original well decreased as expected. In the sample well containing OP50, the fluorescence reading was much higher than for the *GFP* worms alone, indicating that the fluorescence observed from the CLB01 worms was completely overpowered by the OP50.

As the results for the drug screening were to be recorded using the fluorescence readings produced by the CLB01 worms, it was imperative that the OP50 did not interfere with this reading. For this reason it was necessary to identify a concentration of OP50 that did not interfere with the fluorescence reading but still provided enough food for the *C. elegans* worms to remain as healthy as possible for the duration of the screen.

### **6.2.2 Determining the optimum OP50 concentration for compound screening**

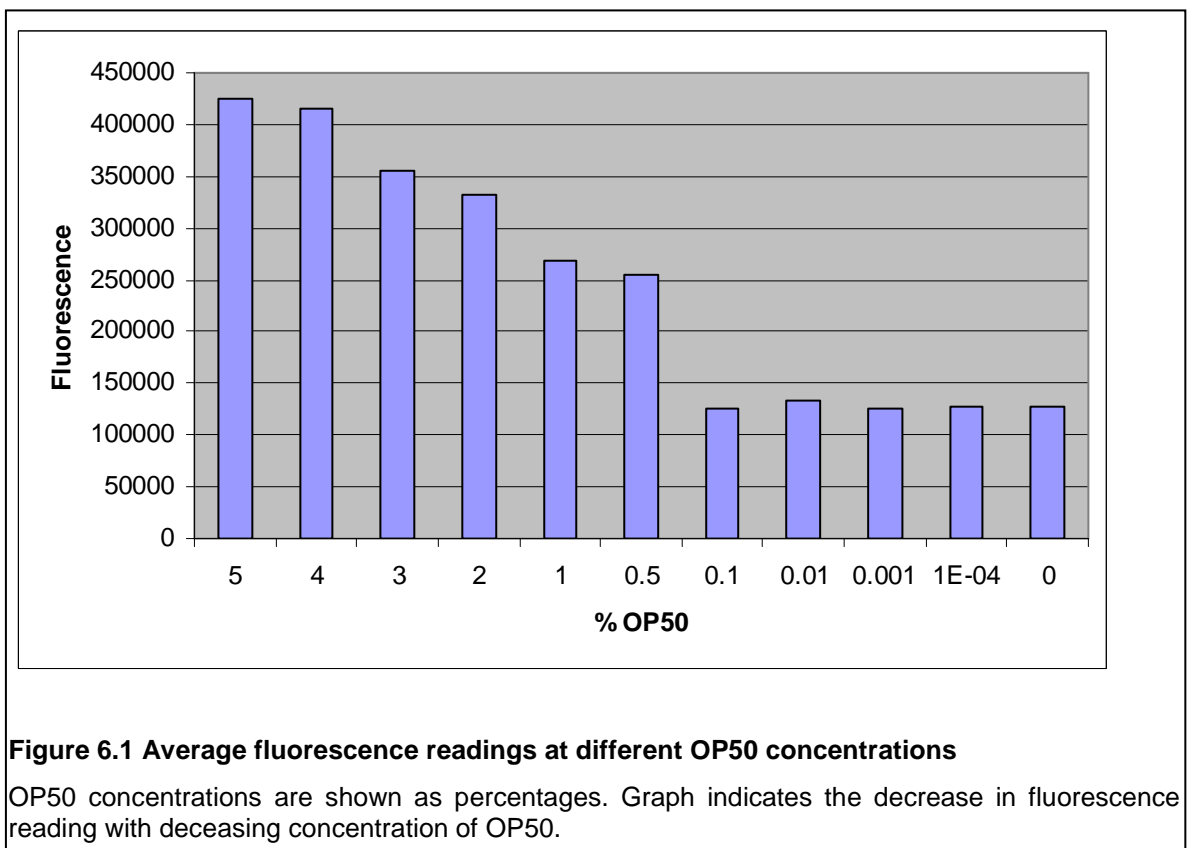
As previously identified, the OP50 bacteria produced a fluorescence reading that interfered with the fluorescence readings from the CLB01 worms. It is not ideal to leave the CLB01 worms in the wells with no food for the duration of the screen thus a workable concentration was required.

A concentration gradient of OP50 was performed to identify the highest concentration of OP50 that could be added to each well without interfering with the fluorescence readings. Ten different concentrations of OP50 overnight culture in 100  $\mu\text{l}$  M9 buffer were analyzed using EnVision and were compared to wells containing M9 buffer with no bacteria added as a control. The concentrations tested were; 5%, 4%, 3%, 2%, 1%, 0.5%, 0.1%, 0.01%, 0.001% and 0.0001%. As expected the fluorescence level decreased with decreasing concentration of OP50 until a point at which it levelled off. This is the baseline fluorescence level and will not interfere with the readings from CLB01 worms.

Fluorescence readings were taken every hour for a period of 18 hours to identify if there was any change in the concentration of OP50 over time. From the results, there did not appear to be any significant change in concentration in

each well over time (results not shown). This confirms that there was no risk of the OP50 concentration increasing during screens and distorting the readings. An average of the readings at each concentration over the 18 hour period was taken to give the level of fluorescence at each concentration (Figure 6.1).

Below a concentration of 0.5% OP50 the fluorescence level appeared to remain fairly constant and the fluorescence reading is the same as that observed in the control wells, containing M9 buffer and no OP50. At these lower concentrations the OP50 does not interfere with CLB01 fluorescence levels and a 0.1% OP50 concentration was therefore selected for use in subsequent screens.

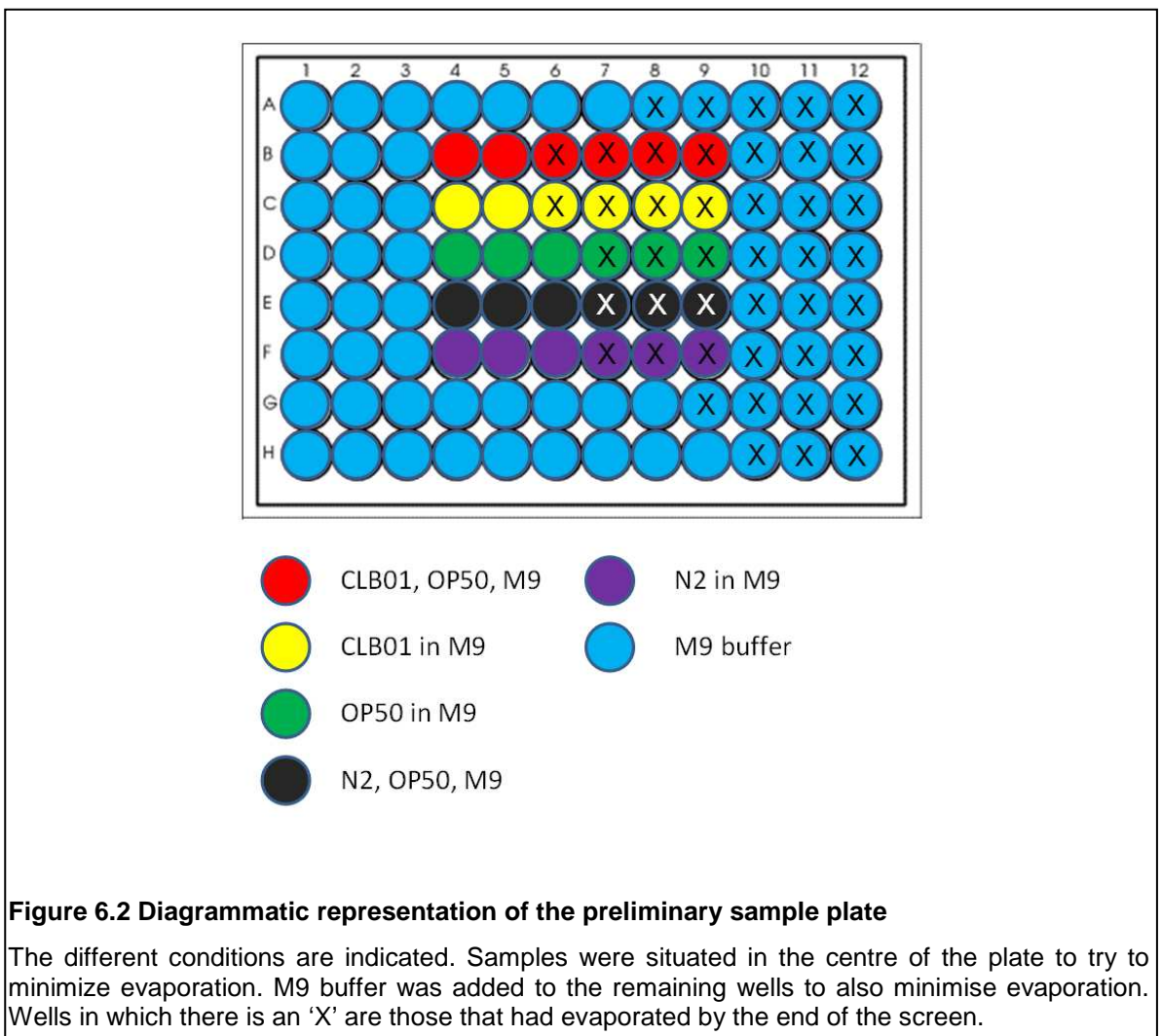


### 6.2.3 A test screen using *C. elegans* to identify optimum conditions

An initial screen of CLB01 worms was set up to run over a period of four days. Fluorescence readings were taken every hour. Six wells were sampled at each condition so as to have enough data to allow for individual well variability and obtain a reliable average. The final volume in each well was 100  $\mu$ l and the concentration of OP50 was 0.1%, conditions identified from the OP50 gradient. The test wells contained: CLB01 worms in M9 buffer and OP50; CLB01 worms in

M9 buffer; OP50 in M9 buffer; N2 worms in M9 buffer and OP50; N2 worms in M9 buffer; and M9 buffer alone (Figure 6.2). In the 96 well plate, the six wells in the middle of each row were used for the samples as it was unknown what the level of evaporation would be throughout the test period. There was expected to be some degree of evaporation, therefore using wells in the centre of the plate and filling the rest of the wells with M9 buffer aimed to combat this.

Taking the precaution to put the samples in the centre of the plate made no difference and evaporation occurred. During the entirety of the screen the plate was left in a darkened chamber, with no lid, inside the EnVision as opposed to it being brought out and the lid replaced after each hourly read. Evaporation was hoped to be at a minimum but on analysis of the plate at the end of the screen, evaporation had occurred in all the wells on the right hand side of the plate, possibly due to an uneven temperature distribution within the EnVision.



Due to the evaporation, six readings per sample were not obtained for the entire screen period. It was possible to determine approximately when the plates dried up looking at the results. For this reason however only results for two or three wells were obtained for analysis of the samples. For subsequent assays, the plates were removed from the Envision and the lids replaced after every reading, thus minimising evaporation.

Upon examination of the worms at the end of the sample period, both those with and without OP50 looked starved. This could be due to incubating the worms for a long period of time (4 days), but even with food the worms may not have been feeding as well as on bacterial seeded plates. There was however still some fluorescence under the GFP microscope at the end of the screen period. The adults showed no fluorescence but there was very faint gut fluorescence in some of the larvae and a number of the larvae also exhibited a degree of hypodermal expression. Although there was a decrease in fluorescence in worms kept under both conditions, fluorescence levels appeared slightly brighter in worms that had been fed on OP50.

As expected the CLB01 worms showed a higher level of fluorescence when compared to wild-type worms. There was no measurable difference between the level of fluorescence in the wells with and without OP50. This indicated that the readings from the CLB01 worms in wells with OP50 were from the worms alone and that there was no background from the OP50. There was a degree of fluorescence expressed from the wild-type worms which was slightly higher than expected. This could be due to the presence of OP50 within the gut of the worms producing some fluorescence and giving a reading. Additionally there were more wild-type worms in the wells than CLB01 worms which could also explain the slight increase in fluorescence. Efforts were made to ensure similar worm numbers in any future samples. From these results it was concluded that wild-type worms produce a fluorescence reading at a low level. Wild-type values can be used as a baseline and anything above this can be taken as the actual level of fluorescence from the CLB01 worms (Figure 6.3).

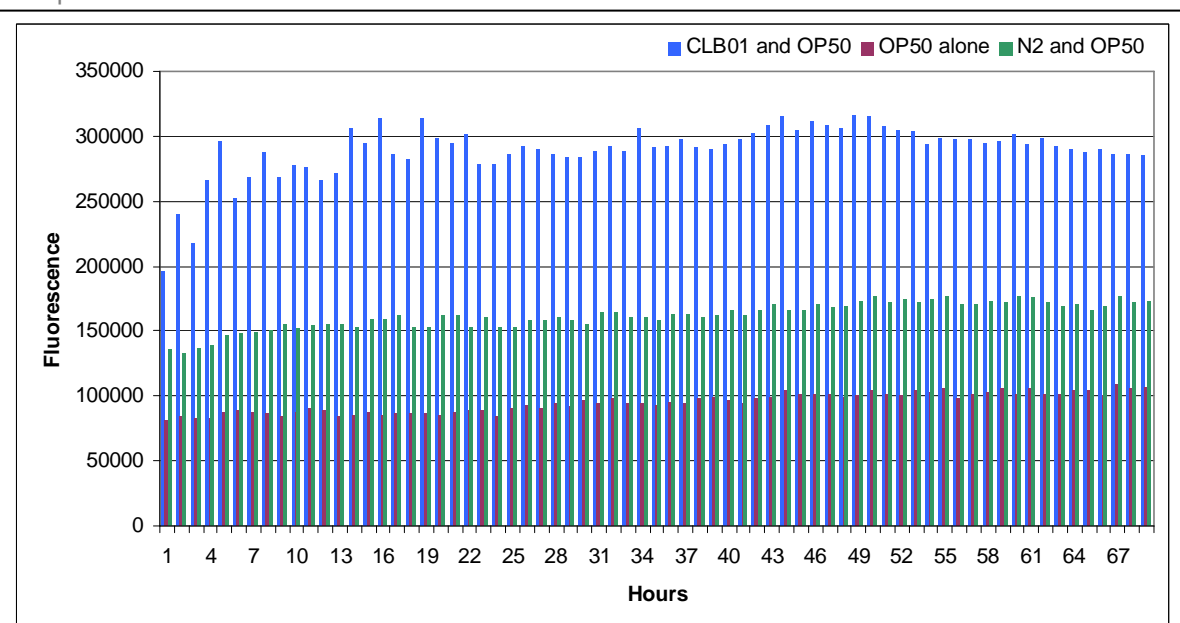
Initially it was presumed that the level of fluorescence would decrease over time, especially if the worms became less viable. This however was not the case

and the level of fluorescence remained fairly constant throughout the screen period of 4 days.

Analysis of the results from the first screen enabled a fluorescence range for each different sample condition and also the average reading to be identified (Table 6.1). The average of the two CLB01 sample conditions without OP50 in the wells gave the higher reads. This is not thought to be significant as there will be a certain degree of variation between worms, which in turn will result in variation between wells.

Looking at the readings from the first test screen, CLB01 worms in OP50 had a minimum reading of 184,937 and a maximum reading of 349,969 (as shown in Table 6.1). The average reading for CLB01 worms in OP50 was 289,354 and was used as the approximate level of fluorescence to aim for in sample wells during the screening of compounds.

The  $Z'$  factor, a measure of assay quality and robustness, was calculated from the mean and standard deviation of the CLB01 and OP50 (positive) and N2 and OP50 (negative) samples.  $Z' = 1 - ((3 \times (\sigma_p + \sigma_n)) \div (\mu_p - \mu_n))$  where  $\sigma$  is standard deviation,  $\mu$  is the mean and p and n are positive and negative controls. The  $Z'$  for the first 24 hours of this assay is 0.52, values of 0.5-0.7 indicate a high quality assay (Gosai *et al.*, 2010).



**Figure 6.3 Fluorescence readings in different sample conditions in a preliminary screen in the absence of compounds**

Graph indicates the difference in the level of fluorescence reading over the 4 day sample period for the three main conditions indicated. n = 2 for each time point for each treatment.

| Sample type     | Fluorescence at 535 nm |         |         |
|-----------------|------------------------|---------|---------|
|                 | Minimum                | Maximum | Average |
| CLB01, OP50, M9 | 184,937                | 349,969 | 289,354 |
| CLB01 in M9     | 244,005                | 352,777 | 303,577 |
| OP50 in M9      | 75,911                 | 105,617 | 94,932  |
| N2, OP50, M9    | 122,561                | 210,671 | 161,933 |
| N2 in M9        | 137,513                | 236,179 | 188,391 |

**Table 6.1 Analysis of the fluorescence levels from the preliminary 4 day test screen**

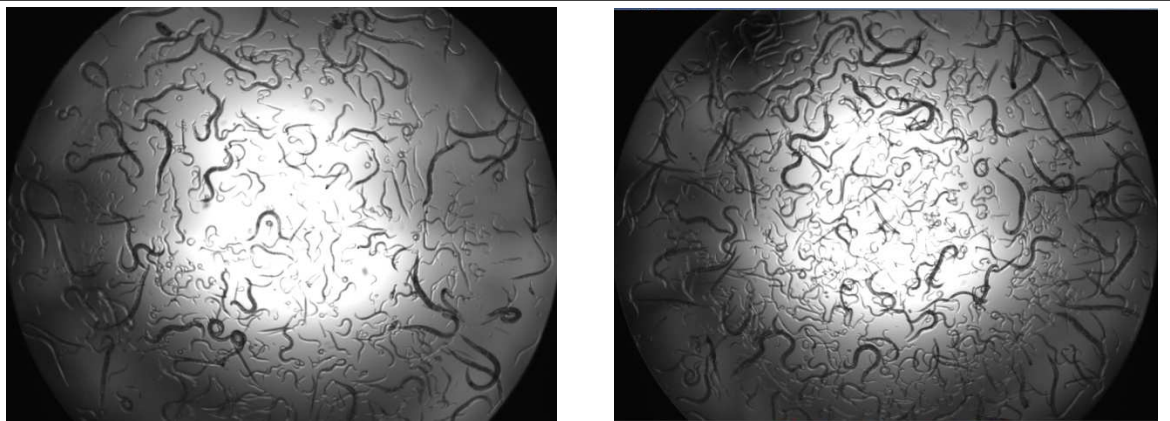
Minimum, maximum and the average reading were taken from the wells over the complete period.

### 6.2.4 Determining the number of CLB01 worms required per sample well

Analysis of worms grown on 10 cm NGM agar plates seeded with OP50 identified that worms left for approximately 6-7 days at 19°C and 60% relative humidity, appear to be at the optimum health for use. At this time point, worms had grown well and were not starved, but there was very little OP50 remaining to contaminate the screen. The worms on one 10 cm agar plate were washed off using M9 buffer and a number of different volumes of worms in M9 buffer were added to different wells to give a variation in worm density and their fluorescence levels measured. Due to fluorescence from OP50, it was necessary to use only clean worms with no OP50 contamination.

One well that had a number of mixed stage worms and no OP50 contamination gave a reading of 429,360, much higher than had previously been observed, but the worm number in the well was large, approximately 200 adult worms. In comparison, a well with slightly less worms but OP50 contamination gave a reading of 406,984 indicating that OP50 contamination had affected this reading. The worms in the clean well were split between two wells and gave readings of 269,254 and 279,435, with similar numbers of worms in each. The adult worm number in each well was counted, showing an average of 68 in the first well and 81 in the second (Figure 6.4). It was therefore aimed to add approximately this number of adult worms to each well. In addition to adult worms, a number of larvae (~100) were present in each well and although these did fluoresce, levels were lower than the adult stage.





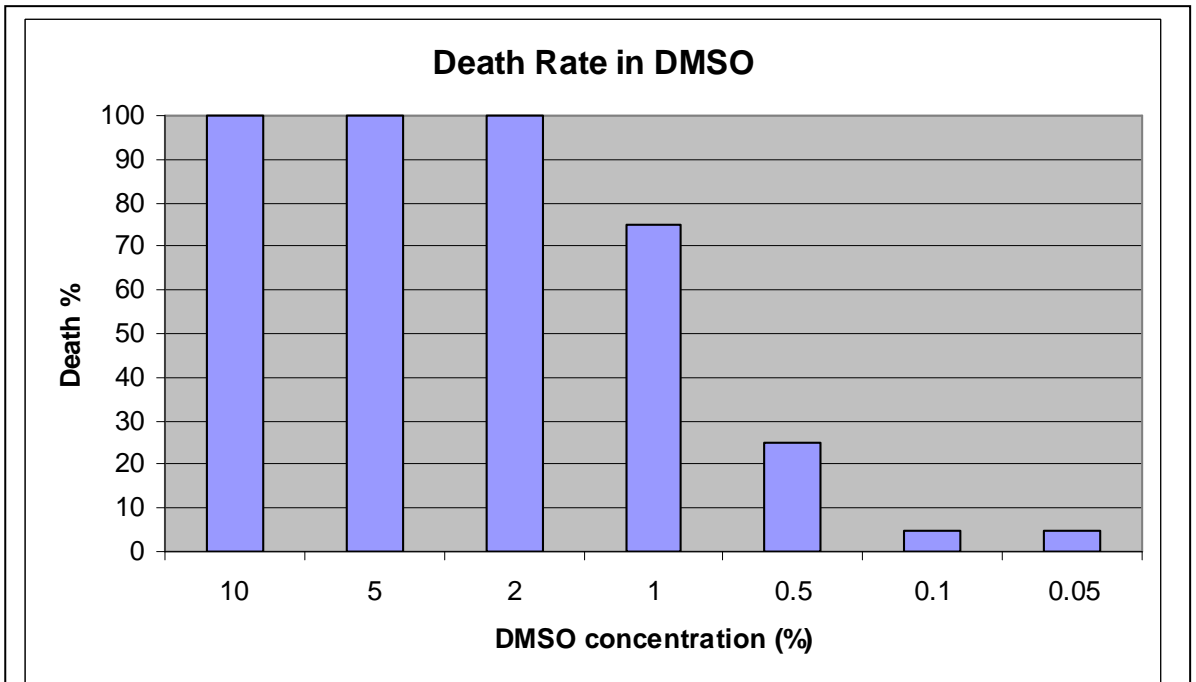
**Figure 6.4 CLB01 worms in sample wells at 2x magnification**

The first image contains approximately 68 adult worms per well and the second 81 adult worms per well. Wells contain all stages of the CLB01 worms.

### **6.2.5 Determining the concentration of DMSO acceptable for screening**

DMSO is stored as a liquid at approximately 100% and is used to solubilise the compounds used for screening. A DMSO curve was set up at seven different concentrations (10%, 5%, 2%, 1%, 0.5%, 0.1% and 0.05%) to determine an appropriate concentration that will not kill the worms or interfere with the results of the screens. Solutions of DMSO were made at 5x the above concentrations and 20  $\mu$ l of each added to 80  $\mu$ l of CLB01 worms, OP50 and M9 buffer, giving a final volume of 100  $\mu$ l. Sample wells containing DMSO at different concentrations were left for 24 hours and the results analyzed at the end of this period.

By manual microscopic examination, there did not appear to be a noticeable difference in the level of fluorescence over the first 24 hour period of exposure to DMSO. Similarly the fluorescence readings from the EnVision supported this observation. For this reason the percentage death in each well was recorded to estimate an acceptable DMSO concentration that could be used for sampling. Figure 6.5 indicates that at high DMSO concentrations (10, 5 and 2%) there was 100% lethality. At 1% DMSO there is 75% lethality, which is too high for sampling. There is observed death at 0.5% but this is at a low level, thus for this reason anything below 0.5% DMSO would be an acceptable concentration to use for sampling. Similarly the 5% lethality observed at 0.1 and 0.05% could be attributed to natural death.



**Figure 6.5 Percentage death of CLB01 worms at different concentrations of DMSO**

Death was estimated from the number of immotile worms identified microscopically after 24 hour exposure to DMSO. n = 4 wells at each concentration.

### 6.2.6 *ELT-2* RNAi as a control to confirm that a decrease in fluorescence will be measurable

The level of fluorescence recorded in the initial 4 day screen using the EnVision did not show a decrease over the test period. Thus, a control was required to confirm that if fluorescence in the worms decreased, this could be identified by the EnVision readings.

*elt-2* RNAi was previously carried out on CLB01 worms by feeding on the *ELT-2* dsRNA expressing bacterial clone, available from the RNAi library (Kamath and Ahringer, 2003). A similar RNAi screen was carried out, but instead of bacterial feeding, the *ELT-2* RNAi bacterial clone was added to the sample wells.

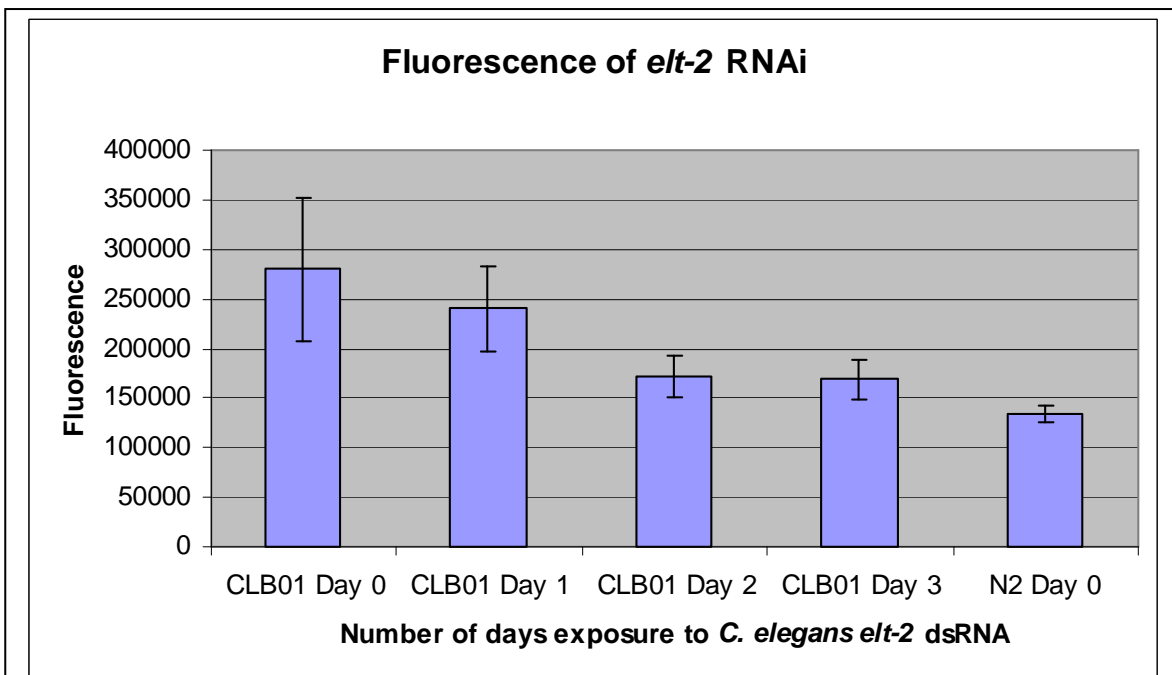
A three day RNAi screen was carried out on CLB01 worms using the same conditions as when fed on OP50 bacteria. Readings were taken every hour and a reduction in fluorescence was expected if the RNAi had been successful. At the end of the test period the fluorescence level in the wells with CLB01 worms fed on *elt-2* dsRNA expressing bacteria had not decreased (results not shown). The reason for this is likely to be because the worms were not feeding well when in

liquid and thus no RNAi effect was observed. Therefore it was examined whether CLB01 worms fed on *elt-2* dsRNA expressing bacteria on agar plates produced a measurable decrease in fluorescence.

A number of plates were seeded with bacteria to obtain a large enough worm number to sample. Initially (Day 0) worms were washed from an OP50 seeded plate, put into wells and the fluorescence level measured to give a baseline. On days one, two and three this was repeated and RNAi fluorescence readings obtained. As readings for worms maintained on plates were not taken automatically, it was not practical to obtain an hourly reading, however taking a reading every day would give an indicator of what reduction can be expected over a period of three days. In all of the sample wells there was a degree of bacterial contamination, especially with the day one sample where there was still some bacteria on the plates when worms were washed off. As with all samples, the worms were washed and the majority of the bacteria removed before sampling with the EnVision.

On each day, six wells containing worms were measured for level of fluorescence. The average value for the six wells of CLB01 worms grown on standard OP50 plates was used as a baseline for maximum fluorescence. Similarly readings for six wells containing wild-type worms were used to indicate the background reading that is observed from wild-type worms.

The results of the RNAi confirmed that a decrease in fluorescence is measurable by the EnVision. The readings on days two and three confirmed that the decrease in fluorescence is significant ( $p < 0.05$ ) compared to day 0. There was a decrease after one day on the RNAi plates but the most significant decrease was between the readings taken on day one and day two. After this time point the readings levelled off and no further decrease was observed (Figure 6.6).



**Figure 6.6** Level of fluorescence of CLB01 worms before and during feeding with bacteria expressing *C. elegans elt-2* dsRNA

Shown are fluorescence levels of N2 worms fed on OP50 as a baseline reading for worms with no GFP fluorescence. Student's t-test used to determine if the results are significant. The readings on Day 2 and 3 confirm that the decrease in fluorescence is significant ( $p < 0.05$ ) compared to Day 0 (mean  $\pm$ STDEV and  $n = 6$ ).

### 6.2.7 Compound selection for use in *C. elegans*

This study was designed to screen for compounds potentially interfering with *C. elegans* ELT-2 GATA transcription factor activity, indicated by a decrease in GFP levels. Therefore compounds which had previously shown detrimental effects on mammalian transcription factor activity were identified in the Pfizer drug database. Compounds are available in either liquid or powder form and can be grouped by selecting for different criteria. The mammalian Inhibitory Concentration resulting in half maximum activity ( $IC_{50}$ ) is one criteria that was used in the selection of compounds for screening. Compounds that are active at workable concentrations were identified. Ideally compounds that are active at concentrations below  $1 \mu M$  were desired as it was an indicator of compound potency. Initial searches did not identify a large number of compounds matching these criteria, thus the search was widened to include compounds that had an  $IC_{50}$  between 1 (classified as high-medium potency) and  $10 \mu M$  (classified as medium-low potency). Overall 594 compounds were selected based on the above criteria.

As the compounds identified showed high potency in mammalian species, there will be no selectivity for *C. elegans*. However potent compounds can be used to carry out searches to find structurally similar compounds that may have the same action, but are less potent for mammalian species.

### 6.2.8 Initial compound screen and analysis of effects

An initial drug screen was carried out on 22 compounds that were identified as being active below 1  $\mu\text{M}$  in an RGate search. Due to the small number of compounds requested they were ordered in powder form and dissolved in DMSO, to produce a stock concentration of 30 mM. One well was sampled for each compound, with any of interest being looked at subsequently in duplicate. Each well consisted of 97  $\mu\text{l}$  of CLB01 worms in M9 buffer, 1  $\mu\text{l}$  diluted OP50 (giving a final 0.1% concentration) and 2  $\mu\text{l}$  of 1/10 dilution of compound (final concentration 60  $\mu\text{M}$ ). One well containing 0.2% DMSO, CLB01 worms, but no compound and one well containing worms in M9 buffer alone were used as controls. The plates remained stacked in the EnVision at room temperature for the duration of the screen and fluorescence readings taken every hour.

Ideally a stock concentration of 10 mM would be obtained for the compounds. This concentration would ensure that the compounds were fully dissolved. However due to the low tolerance of *C. elegans* to DMSO a higher stock concentration of compound is required to ensure as high a sample concentration of compound as possible. This high stock concentration resulted in a number of compounds coming out of solution, looking cloudy or being very viscous, which was unavoidable.

For this first screen, fluorescence was measured in the wells containing CLB01 worms before the compounds were added and showed fluorescence readings within the desired range (results not shown). After addition of the compounds, a number of the wells showed a very high level of fluorescence; this was presumed to be due to fluorescence from the compounds themselves. Readings were taken of the compounds in M9 buffer alone to confirm that some compounds did fluoresce. The fluorescent nature of some of the compounds was not ideal, however those compounds that did fluoresce could be easily identified both by the high fluorescence reading from the EnVision and visually with a GFP

microscope (Table 6.2). The level of fluorescence of the compounds in M9 buffer without any worms was recorded every hour over a 48 hour period, and found not to change over this time. This ensured that any change in fluorescence would be specific to the worms rather than the compound. For compounds that fluoresce highly, initial readings were taken into consideration when analyzing the results.

| Compound  | Compound fluorescence alone | Fluorescence observed visually | Compound & worm fluorescence |
|-----------|-----------------------------|--------------------------------|------------------------------|
| 1/1       | 78,883                      | N                              | 300,166                      |
| 1/2       | 84,575                      | N                              | 289,881                      |
| 1/3       | 146,833                     | N                              | 224,950                      |
| 1/4       | 261,845                     | Y                              | 309,226                      |
| 1/5       | 5187,040                    | Y                              | 6515,069                     |
| 1/6       | 374,127                     | N                              | 504,552                      |
| 1/7       | 1,090,288                   | Y                              | 1,005,947                    |
| 1/8       | 82,901                      | N                              | 220,632                      |
| 1/9       | 170,919                     | N                              | 357,594                      |
| 1/10      | 519,384                     | Y                              | 500,586                      |
| 1/11      | 78,256                      | N                              | 172,914                      |
| 1/12      | 73,873                      | N                              | 183,003                      |
| 1/13      | 78,458                      | N                              | 308,301                      |
| 1/14      | 1,003,763                   | Y                              | 578,885                      |
| 1/15      | 451,747                     | Y                              | 498,798                      |
| 1/16      | 114,333                     | N                              | 249,902                      |
| 1/17      | 78,185                      | N                              | 289,399                      |
| 1/18      | 139,365                     | N                              | 197,360                      |
| 1/19      | 102,940                     | N                              | 188,531                      |
| 1/20      | 160,637                     | Y                              | 474,812                      |
| 1/21      | 90,842                      | N                              | 226,095                      |
| 1/22      | 68,437                      | N                              | 245,506                      |
| 0.2% DMSO | 86,318                      | N                              | 166,331                      |
| M9 buffer | 84,516                      | N                              | 192,504                      |

**Table 6.2 Analysis of the 22 compounds used in the first compound screen**

Values indicated for the compounds alone and with worms at 0 hours (measured by the EnVision), compared to visual analysis. Fluorescence levels and visual analysis correlate in all compounds except for compound 1/20.

Initial analysis after 24 hours indicated that for most wells, fluorescence levels remained fairly constant, although a few wells showed a slight decrease. Compound 1/20 showed a slightly larger decrease over the first 24 hours. This was one of the compounds showing visual fluorescence however, the reading from the compound alone was quite low (160,637). The first reading for CLB01

worms in this compound was 474,812, decreasing to 222,860 at 24 hours. This finding may or may not be important, but is worth noting at the 24 hour mark.

After 48 hours of worm exposure to the compounds, there were a number of wells in which the worms showed a decrease in fluorescence by eye. Some decrease is to be expected due to the worms not being as healthy as they were at the beginning of the screen. To try and identify the potentially active compounds, combined analysis of the fluorescence readings and visual analysis of the worms microscopically would have to be taken in to consideration. Those compounds that showed a decrease in the level of fluorescence using the EnVision over the 48 hour screen period were compared to those wells in which the level of fluorescence identified using a fluorescence microscope decreased and additionally there was a significant level of death. One problem that distorts the results provided by the EnVision is the auto-fluorescence that occurs in some worms when they die. This auto-fluorescence appears to be restricted to adult worms and means that even if a compound does not show a decrease in the level of fluorescence recorded by the EnVision or shows a decrease and then an increase it may still be active as there may be auto-fluorescence. It is for this reason that visual analysis was also carried out. Conversely, during the DMSO screen there was no observed auto-fluorescence but worms died and continued to produce GFP fluorescence.

Analysis of the results at the end of the first screen identified a few common factors. The 24 hour time point appeared to be important, as it was thought that by this point the largest decrease in fluorescence should have occurred. This presumption was made from analysis of the RNAi screen and identifying when a decrease in fluorescence was first observed (Figure 6.6). For a number of the compounds there is either a gradual decrease up until this point or a sharp decrease just before it. Post 24 hours the levels either increased, decreased or remained constant. Wells showing an increase were thought to be due to an increase in auto-fluorescence from the dead worms. Those that decreased were most likely due to a continuing drop in GFP expression and those that remained constant probably reflect both auto-fluorescence and GFP expression, these compounds may still target GFP expression however, GFP effects may be masked. A number of the compounds (1/7, 1/10, 1/15, 1/18 and 1/22) appeared to have no effect on the level of fluorescence and the early readings from the



EnVision remained fairly constant. Even if these compounds kill the worms, they are not affecting GFP levels. Compounds that appeared to be most effective at killing the worms include 1/7, 1/13, 1/15, 1/20, 1/21 and 1/22. Compounds 1/7, 1/15 and 1/22 are included in both categories indicating that death does not appear to be related to loss of GFP expression. Compounds 1/13, 1/20 and 1/21 were selected for further analysis as they have been identified as compounds that are effective at killing worms.

Each of the three compounds demonstrated some decrease in fluorescence using EnVision over the first 24 hours suggesting an effect on gut GFP expression. Compound 1/13 did not show an obvious decrease in fluorescence level by eye. Compound 1/21 showed reduced fluorescence but this was not as convincing as the effects observed by compound 1/20. Therefore only compound 1/20 was examined in more detail.

### **6.2.9 Large scale screening of compounds in *C. elegans***

From this first screen there was a large volume of data gathered. This was manageable due to the small sample number. However, a quicker and more efficient method of identifying potentially active compounds was required for larger screens. Fluorescence readings for the initial screen were taken every hour which proved to be unnecessary and readings were therefore reduced to every two hours for subsequent screens. Plates remained stacked in the EnVision for the period of the screen and the lids replaced after each 2 hourly read. Only readings at 0, 24 and 48 hours were used to identify a decrease and those compounds that looked of interest at these three time points were analyzed in more detail over the full 48 hour period. Effects of these compounds were identified by their fluorescence reading in the EnVision, fluorescence by eye and also ability to kill the worms. Any compounds that were of interest in all cases could be looked at in more detail.

The majority of the compounds used for large-scale screening were ordered as liquids due to the high number requested. There are a few disadvantages when ordering compounds as liquids, one being the significantly lower concentration that can be used for screening, due to constraints on the DMSO concentration which can be used. The liquids are supplied at a concentration of 4 mM in 100%

DMSO. To achieve a workable concentration of DMSO a volume of 0.5  $\mu\text{l}$  of the compound in DMSO was added to give a final sample volume of 100  $\mu\text{l}$ . Using the calculation

$$(IC)(IV) = (FC)(FV)$$

(initial concentration)(initial volume) = (final concentration)(final volume) results in a final compound concentration of 20  $\mu\text{M}$  in each sample well. This is much lower than the 60  $\mu\text{M}$  that can be achieved when using compounds from powder form. Secondly, a much lower compound volume is supplied, meaning less is available for further screening of compounds of interest.

Combining the results of the initial screen and the large scale screen, of the 594 compounds examined, 50% of these showed a decrease in fluorescence at the 24 hour time point. After this, only 47% of these 50% were lower than the 0 hour reading at 48 hours. This was a very large number of compounds that could be potentially affecting fluorescence levels and for this reason the other methods of analysis (visual GFP observation and percentage death) were taken into consideration when selecting compounds for further analysis. Compounds that met all three criteria were identified and dose response curves generated. One of these, compound 1/20, was identified as potentially affecting gut gene expression in *C. elegans* and was therefore selected to study in more detail.

### **6.2.10 Searching for compounds similar to those of interest**

Compound 1/20 that was identified as being of interest in the first screen was examined in more detail in RGate. A search was carried out to identify compounds with a similar structure, likely to have similar activity. Most drugs that work by the same mechanism of action are highly similar in their structural composition (McCracken and Lipkowitz, 1990) for this reason if compound 1/20 is acting on the desired target, compounds with a similar structure should also follow this pattern. A 90% structural similarity was used as the cut-off point. Compounds within the criteria were identified using the RGate computer programme, using parameters such as the number and position of different elements. If the activity observed by the first compound was real then these compounds should also have an effect. Ideally a Structure Activity Relationship

(SAR) should be observed, as adding or removing groups from the structure should have an effect on the potency or even render the compound inactive. 37 compounds were identified as being structurally similar to compound 1/20 and ordered as powders, however only 31 were available for sampling. Compound 1/20 was also added to this screen, but at an unknown location.

### **6.2.11 Creating a dose response curve for compounds of interest**

Dose response curves were generated for all compounds of interest to determine a linear relationship and threshold level of drug. For those compounds that were supplied as powders the concentrations used for the curve were 60  $\mu\text{M}$ , 20  $\mu\text{M}$ , 6  $\mu\text{M}$ , 2  $\mu\text{M}$ , 0.6  $\mu\text{M}$  and 0.2  $\mu\text{M}$ . When creating the curve, a 50x working concentration was made at each dilution, each well contained 97  $\mu\text{l}$  of CLB01 worms in M9 buffer, 2  $\mu\text{l}$  of compound and 1  $\mu\text{l}$  of OP50.

For those compounds supplied as liquids, lower concentrations were used (20  $\mu\text{M}$ , 5  $\mu\text{M}$ , 2  $\mu\text{M}$ , 0.5  $\mu\text{M}$ , 0.2  $\mu\text{M}$  and 0.05  $\mu\text{M}$ ) and each well contained 98.5  $\mu\text{l}$  of CLB01 worms in M9 buffer, 0.5  $\mu\text{l}$  of compound and 1  $\mu\text{l}$  of OP50.

Dose response curves were calculated using the fluorescence readings at 24 and 48 hours and expressing these as a percentage reduction of the 0 hour reading at different concentrations of each compound. Ideally a decrease in the percentage reduction as the concentration of the compound decreases should be observed. Values observed at the 48 hour point should be lower than those observed at 24 hours.

### **6.2.12 Screening and analysis of compounds of interest**

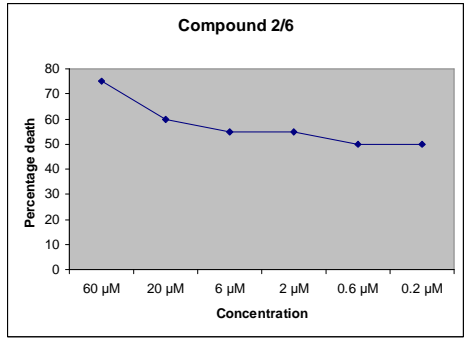
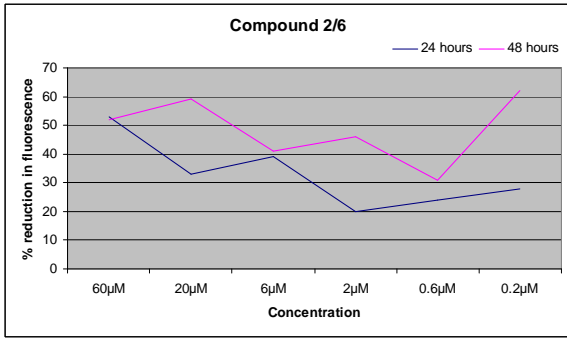
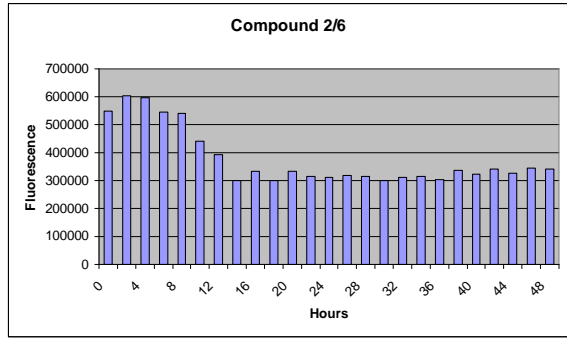
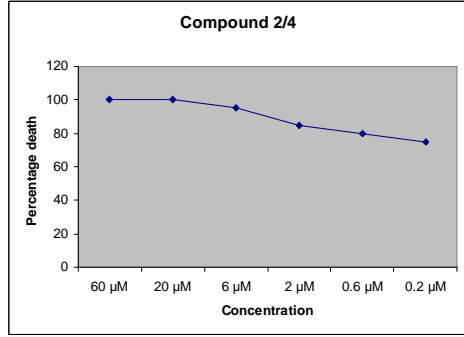
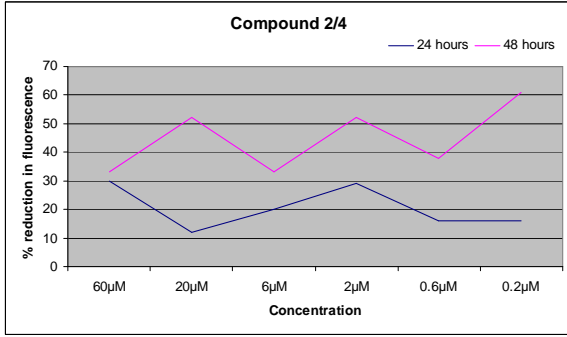
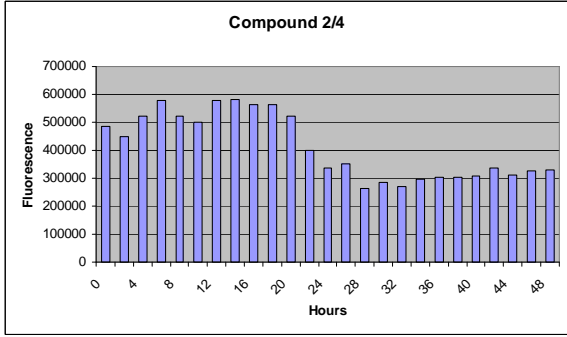
The compounds identified in the similarity search were screened and the results analysed. A number of the compounds showed a decrease in fluorescence reading over the period of the screen; compounds 2/2, 2/3, 2/4, 2/6, 2/7, 2/9, 2/10, 2/12, 2/13, 2/15, 2/16, 2/17, 2/18, 2/20, 2/23 and 2/24 all had notable decreases in the fluorescence readings from the EnVision. The largest decrease was over the first 24 hours; after this point there was an increase in the reading in a number of the compounds. A slight increase is not thought to be of interest. On visual analysis there appeared to be a number of compounds that resulted in

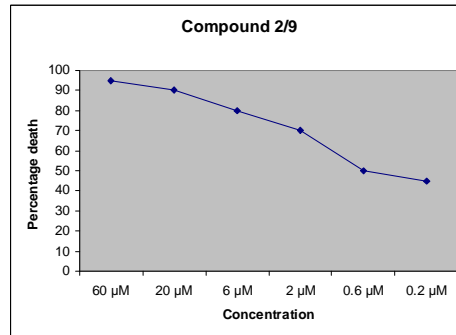
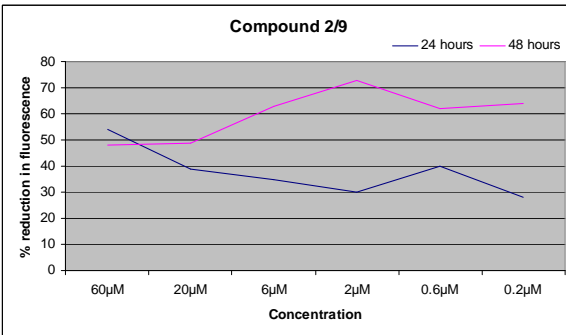
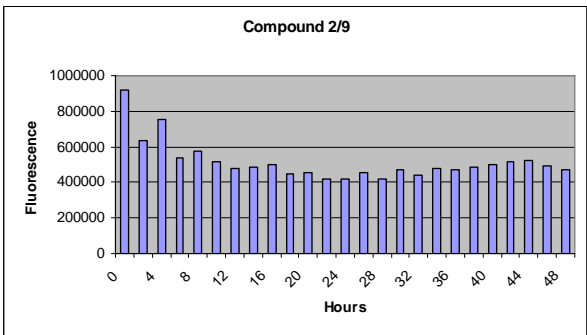
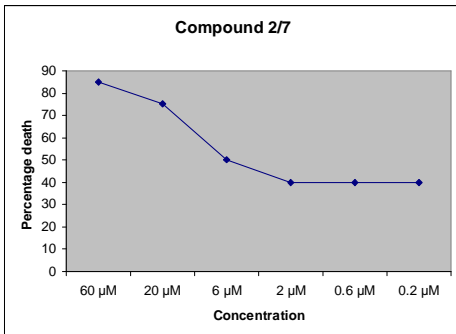
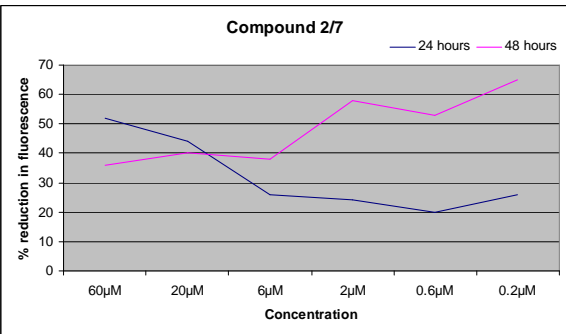
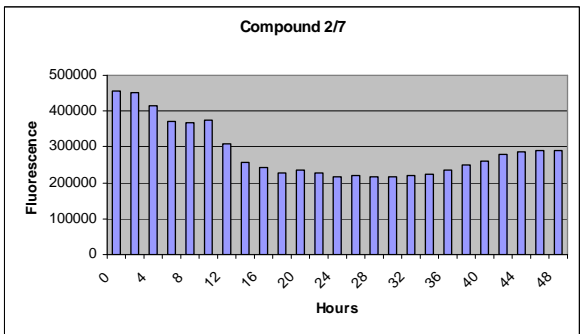
lethality in addition to a decrease in fluorescence. These were; 2/4, 2/6, 2/7, 2/8, 2/9, 2/10, 2/14, 2/27, 2/31 and 2/32. The fluorescence levels for those compounds which appear in both results (2/4, 2/6, 2/7, 2/9 and 2/10) are indicated in Figure 6.7. From the fluorescence data, compounds 2/4, 2/8 and 2/9 closely follow the desired pattern of a decrease in fluorescence at 24 hours and also at 48 hours. Compound 2/7, although resulting in a slight increase in fluorescence reading at 48 hours, was selected for closer analysis as it looked to be effective at killing *C. elegans* worms. Compound 2/10 was interesting as the fluorescence reading throughout the screen period remained almost constant, however there was a higher initial reading which is why a decrease in fluorescence was identified. Observations by eye identified fluorescence from compound 2/10 and also auto-fluorescence in the wells which may have interfered with the readings.

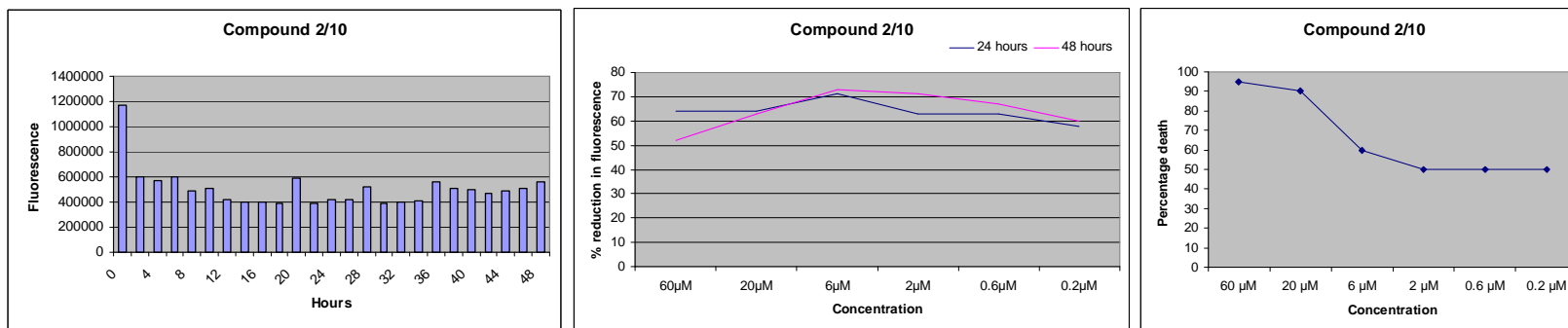
These compounds were used to create curves and determine a dose response. Compound 2/7 was identified by both the EnVision and by eye as having activity and it was later revealed as being the same compound as number 1/20 from the first screen, confirming that the effects observed with this compound in the previous screen were repeatable. In addition to the fluorescence reading and dose response curves, the percentage death in each of the wells at each concentration was recorded. If the effects were real, and a true dose response is being observed, the death rate should be higher at higher concentrations (Figure 6.7).

Examination of the dose response curves in more detail showed that compounds 2/4, 2/6 and 2/9 look to be quite similar. As a general trend for compound 2/4, the GFP fluorescence level decreases with increasing drug concentration. However, this compound shows auto-fluorescence, which is greater at higher concentrations. This could account for a lower percentage decrease of fluorescence at higher drug concentrations. The only compound to follow the desired visual pattern of a linear relationship between fluorescence level and drug concentration was compound 2/7 at 24 hours. Additionally, observations by eye support the readings observed by the EnVision and relationship between drug concentration and fluorescence.

All of the compounds exhibited a high percentage of death at the highest screening concentration, suggesting non-specific toxicity. Only for compound 2/4 was a high level of death observed, even at low concentrations (Figure 6.7). Visual analysis identified this compound as being effective at killing worms of all stages.







**Figure 6.7 Graphic analysis of the compounds of interest**

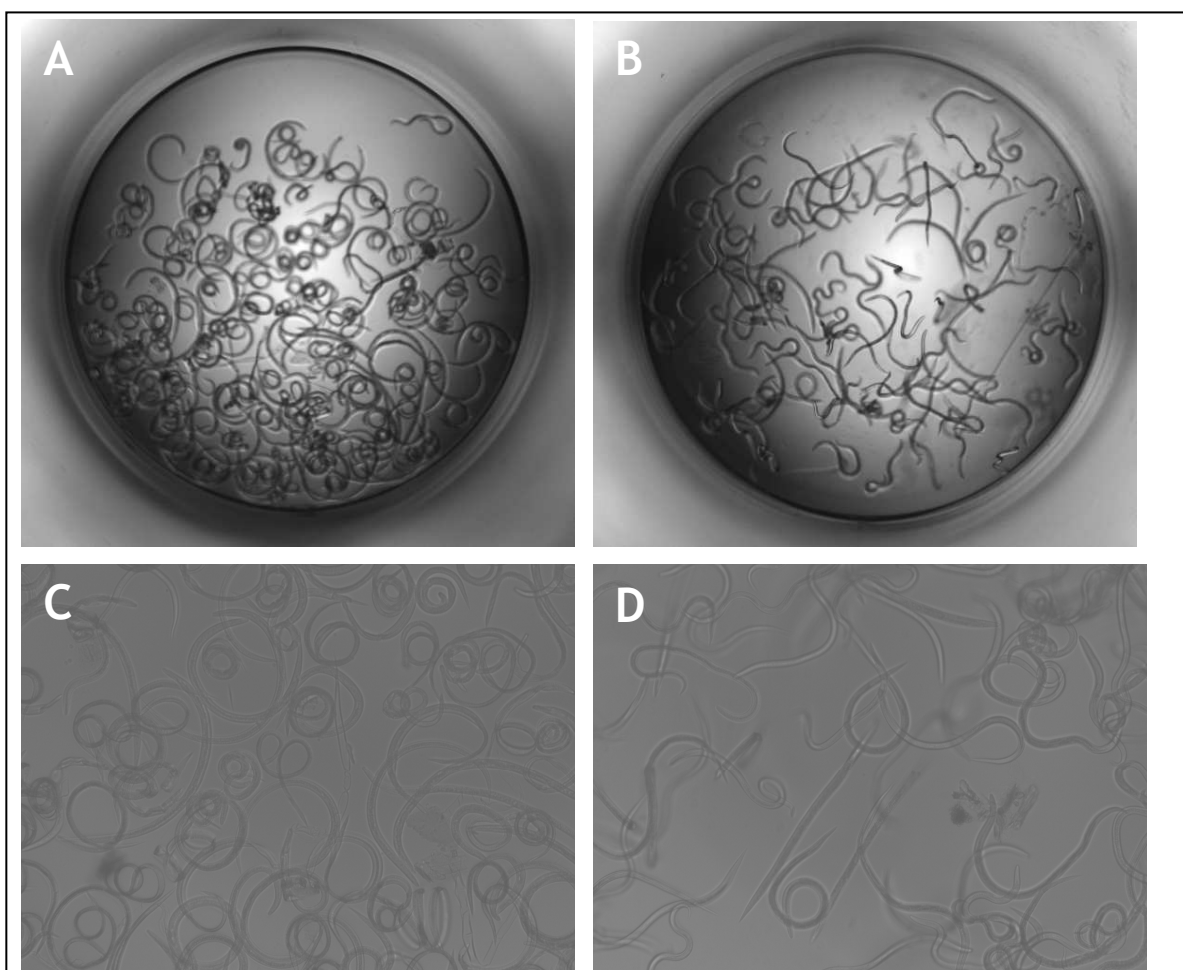
Graphs indicate the decrease in fluorescence readings for each compound throughout the 48 hour screen period when screened at 60  $\mu$ M. Dose response curves for the compounds of interest indicate the % reduction in GFP fluorescence at 24 and 48 hours, relative to the 0 hour value for each compound. Percentage of death of the worms in each compound after 48 hours are shown, with values being the average of visual counting of 2 wells at each concentration. Although the compounds shown resulted in a decrease in fluorescence over time, only compound 2/7 showed a desired dose-response curve (at 24 h).



### 6.2.13 Screening of compounds using *H. contortus* L3 larvae

All of the compounds screened in *C. elegans* were additionally screened in *H. contortus*. The *H. contortus* larvae were recovered from American sheep fed on pasture and thus parasite resistance status reflects that of the natural endemic population (Section 2.2.2). Compounds were added to produce a final concentration of 80  $\mu\text{M}$  per well, with each compound being examined only once. The exsheathed larvae were observed by eye at 24 hours and after 4 days, and positive results were recorded. The method of *H. contortus* larval preparation was the standard procedure developed at and followed by Pfizer employees.

Only three compounds showed any activity in *H. contortus*; these were termed compounds X, Y and Z. 'Active' compounds were identified by the visual observation of wells and recording those in which a large number of *H. contortus* larvae were dead or dying (Figure 6.8). Structurally, compounds X and Y were almost identical. In the *C. elegans* screens, these compounds had not been identified as having an effect on GFP levels, however the results of the *C. elegans* screens for these three compounds were subsequently re-examined in more detail to confirm observations.



**Figure 6.8** *H. contortus* L3 larvae in active and inactive compounds

These images show the contrast between the larvae that are dead and curled up (A and C) and those that are still alive and uncurled (B and D), taken on day 4 of the screen. A and B at x2 magnification and C and D at x10 magnification.

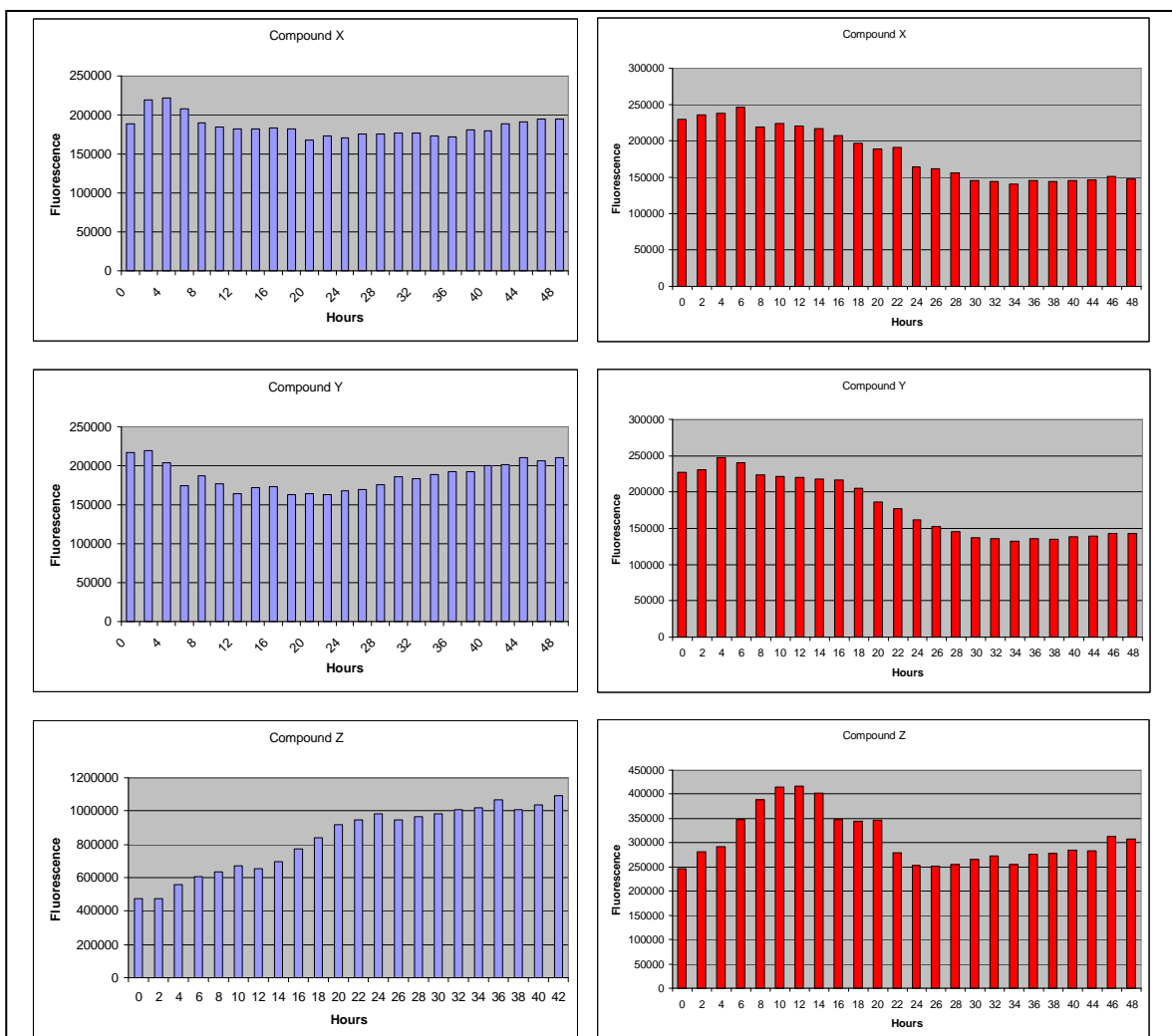
Upon re-examination of the results from the initial *C. elegans* screens, the compounds identified as being active in the *H. contortus* screen, resulted in death of *C. elegans* worms at the 24 hour time point. Compounds X and Y appear to be killing all stages of *C. elegans*. Compound Z appears to be more effective on adults, with a very low rate of larval killing. Visual examination showed approximately 50% death at 24 hours in compounds X and Y and approximately 75% death with compound Z. Although at the 24 hour stage, compound Z appears to be most effective at killing the worms, there does not appear to be a visual decrease in GFP level. In the presence of compounds X and Y, worms showed a slight decrease in level of fluorescence after 24 hours.

After 48 hours in the compounds, the percentage death of *C. elegans* are very similar to those observed at the 24 hour mark. There was a slight increase in killing with compound Y, increasing up to 60%, however this value was still relatively low. Compound Z, which appeared to be most effective at killing *C. elegans*, did not appear to have any effect on GFP at 48 hours, while slight decreases were observed with X and Y.

In the first *C. elegans* screens, a decrease in fluorescence of the compounds active in *H. contortus* screens had not been noted. To re-test this, three sample wells were set up for each compound and an average taken of the fluorescence readings. For compounds X and Y, after the re-examination screen, the results indicated a decrease in fluorescence at both the 24 and 48 hour time points (Figure 6.9), in contrast to the initial screen in which an initial decrease was followed by an increase. For compound Z however, the readings over the 48h screen period were very varied and did not appear to follow any pattern. The results from the re-examination screening of compounds X and Y look to be more reliable, however for compound Z, the initial results looked to be more accurate.

The 24 and 48 hour readings for compounds X and Y demonstrated a decrease in fluorescence at both time periods (Figure 6.9). This differed from the readings in the initial screen in which the readings had decreased slightly by 24 hours but had increased again by 48 hours. As a number of wells were set up for each compound, the significance of the observed results was calculated using the t-test to determine p value (<http://studentsttest.com/>). Calculations confirmed

that the values for compounds X and Y in the second screen are significant to a 99% confidence in the results ( $p < 0.001$ ). As the values increased for compound Z the results are not significant. Therefore, re-examination screening suggests that compounds X and Y share detrimental effects on both *H. contortus* and *C. elegans* and this is affecting *C. elegans cpl-1* GFP levels.



**Figure 6.9** Fluorescence readings of the compounds identified by the *H. contortus* screen

Original graphs from the first *C. elegans* screen are in blue and the average values of the re-examination screen in red. Original graphs did not show a significant decrease in fluorescence throughout the screen period. However in the repeat screen fluorescence levels for compounds X and Y decreased throughout the time period.

### 6.2.14 Further analysis of two of the Pfizer compounds

Two of the compounds (1/20 and 2/4) identified from the larger scale screen at Pfizer as being of interest in *C. elegans*, were analysed to examine their effects in more detail. Screening was carried out in 96 well plates with a final compound concentration of 60  $\mu\text{M}$ . L2 and L3 stage CLB01 worms were added to the wells and those exposed to both compounds displayed a decrease in GFP and

developed more slowly compared to DMSO control wells. This is consistent with *elt-2* RNAi effects, in which if young larvae are put on the plates they are unable to develop properly.

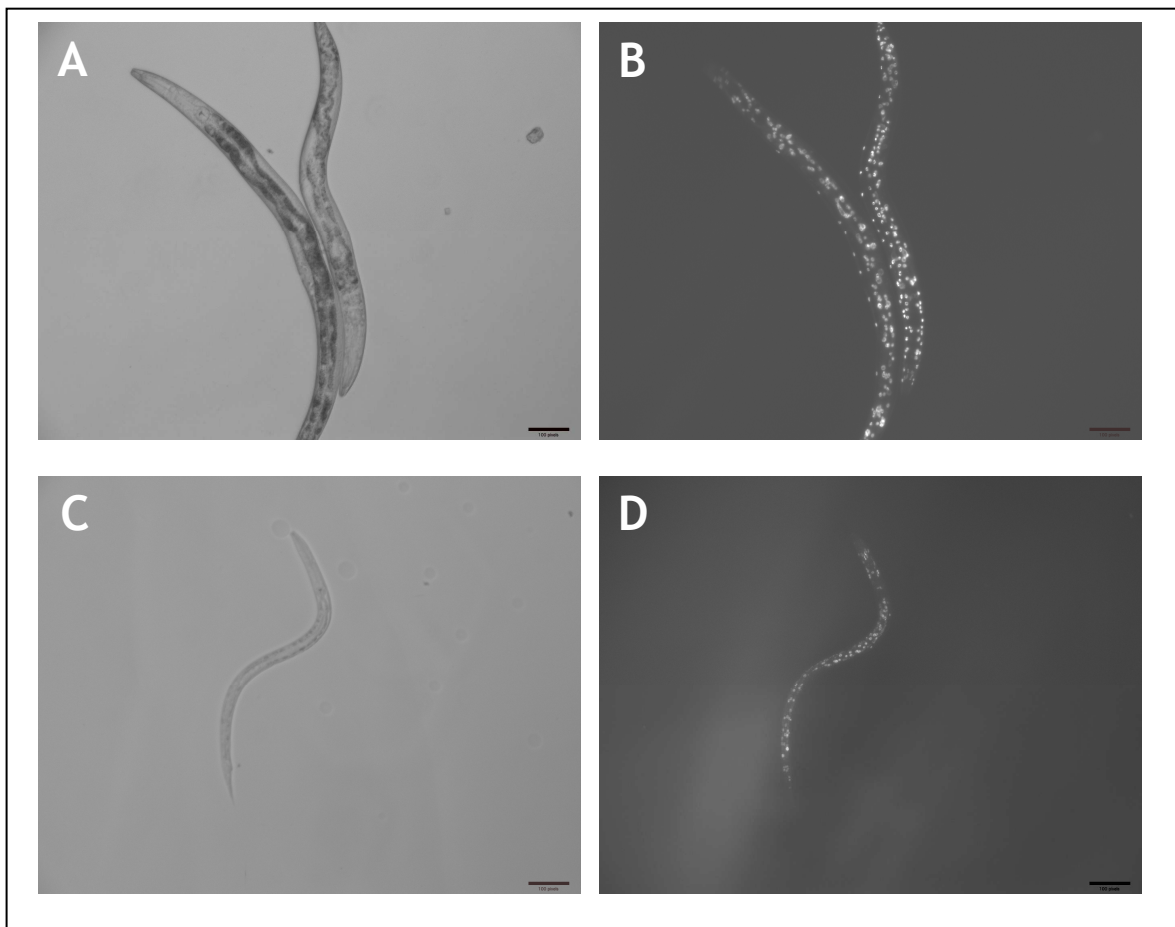
CLB01 larvae exposed to both compounds for two days and control worms in DMSO were observed at high power to assess development. The control worms were larger and looked as if they had been feeding, with GFP fluorescence visible in both gut and hypodermal cells. Worms in the wells containing the compounds were visibly smaller but there was still GFP fluorescence (Figure 6.10).

Addition of older, L4/young adult, CLB01 worms to the wells resulted in a decrease in GFP fluorescence. Moreover, it was obvious that fewer progeny were developing from adults exposed to the compound compared to DMSO control wells. To confirm that the compound is having a specific effect on gut gene fluorescence rather than a generalised reduction of gene expression a SX328 (*myo-2::gfp*) worm strain was used as a control. This strain is GFP positive in the pharynx and after exposure to the compounds there was no observed decrease in GFP, however the worms appeared less healthy compared to the control wells and showed the same embryonic effect as the CLB01 strain (results not shown). To examine the embryonic effect the compounds were having on progeny production, eight L4 CLB01 worms were picked into compound or DMSO control wells and left to develop for three days. This was carried out in triplicate and on day three the L1 progeny in the wells were counted. The results of this are indicated in Table 6.3; it was observed that over three times as many progeny are produced from worms in control wells.

|          | Control | Compound 1/20 | Compound 2/4 |
|----------|---------|---------------|--------------|
| Sample 1 | 214     | 108           | 26           |
| Sample 2 | 214     | 24            | 46           |
| Sample 3 | 251     | 68            | 74           |
| Average  | 226     | 67            | 49           |
| STDEV    | + 21.36 | + 42.01       | + 24.11      |

**Table 6.3 Number of progeny in compound and control wells after three days**

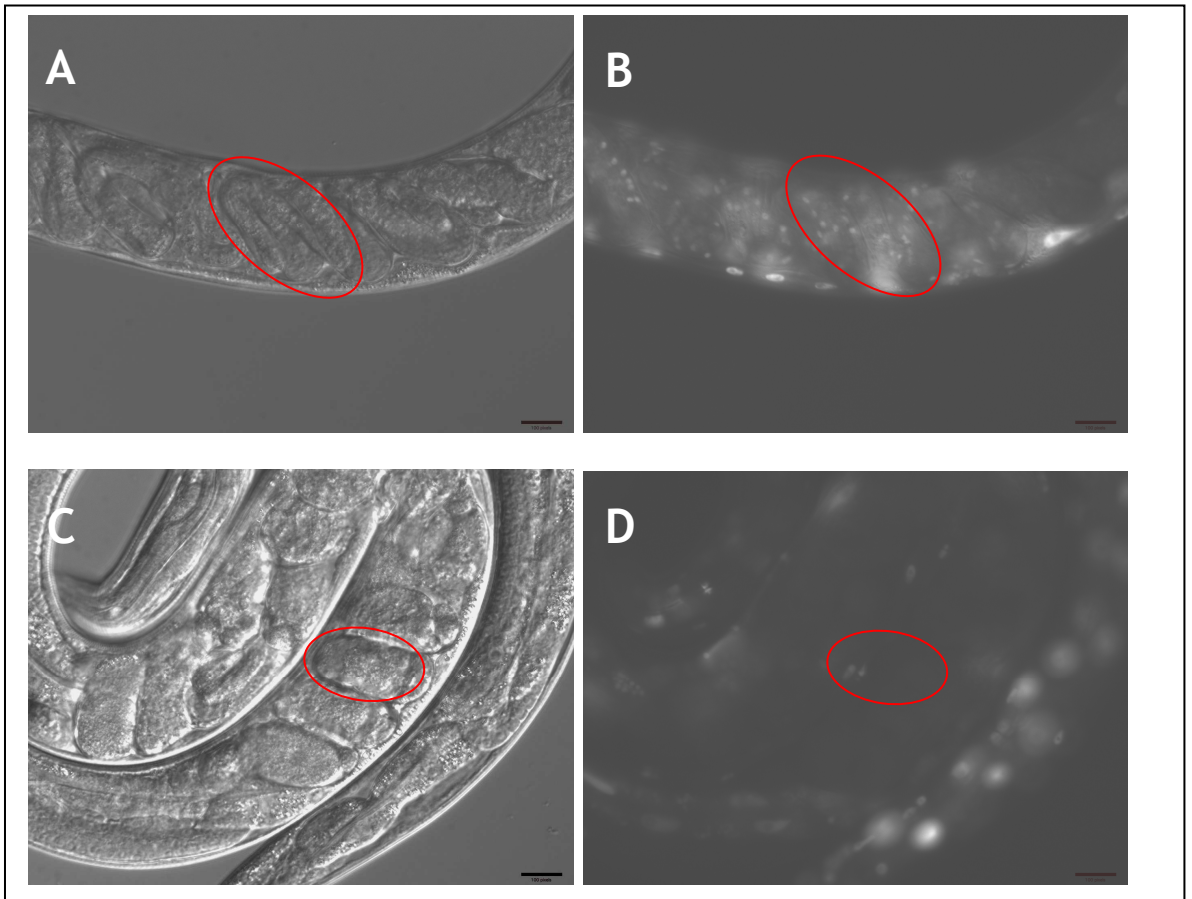
8 L4 larvae added per sample well and the progeny counted after 3 days. The statistical significance of compound 1/20 and compound 2/4 was determined by Student's t-test which was  $p < 0.005$  and  $p < 0.001$  respectively.



**Figure 6.10 Slower development of CLB01 larvae following exposure to compound**

(A and B) CLB01 worms 3 days after the addition of L3 larvae to wells containing DMSO. Normal development to adult and GFP fluorescence in gut and hypodermal cells observed. (C and D) CLB01 worms 3 days after the addition of L3 larvae to wells containing compound 1/20. Slower development and reduced GFP fluorescence compared to control worms. All images at x10 magnification.

It was of interest to determine at what point embryonic development is being interrupted and therefore adult CLB01 worms from both the control and compound wells were observed at high power. As the worms were being grown in liquid culture it is normal for embryos to develop further within the worms and thus in some worms fully elongated embryos could be observed. Egg development in the compound wells did not look normal, with a pitted or vacuolar appearance in some of the embryos. Development appeared to be most affected in younger embryos that started developing when the adult worms had been exposed to the compounds for a longer time. Figure 6.11 indicates the difference in development of the embryos, with those in the control wells developing further than those exposed to compound. GFP fluorescence is clearly visible in the control embryos however this is decreased and hardly visible in compound wells. A few embryos in the compound wells, particularly those produced first, showed normal, full development (Figure 6.11 palette D) but GFP fluorescence was barely visible.



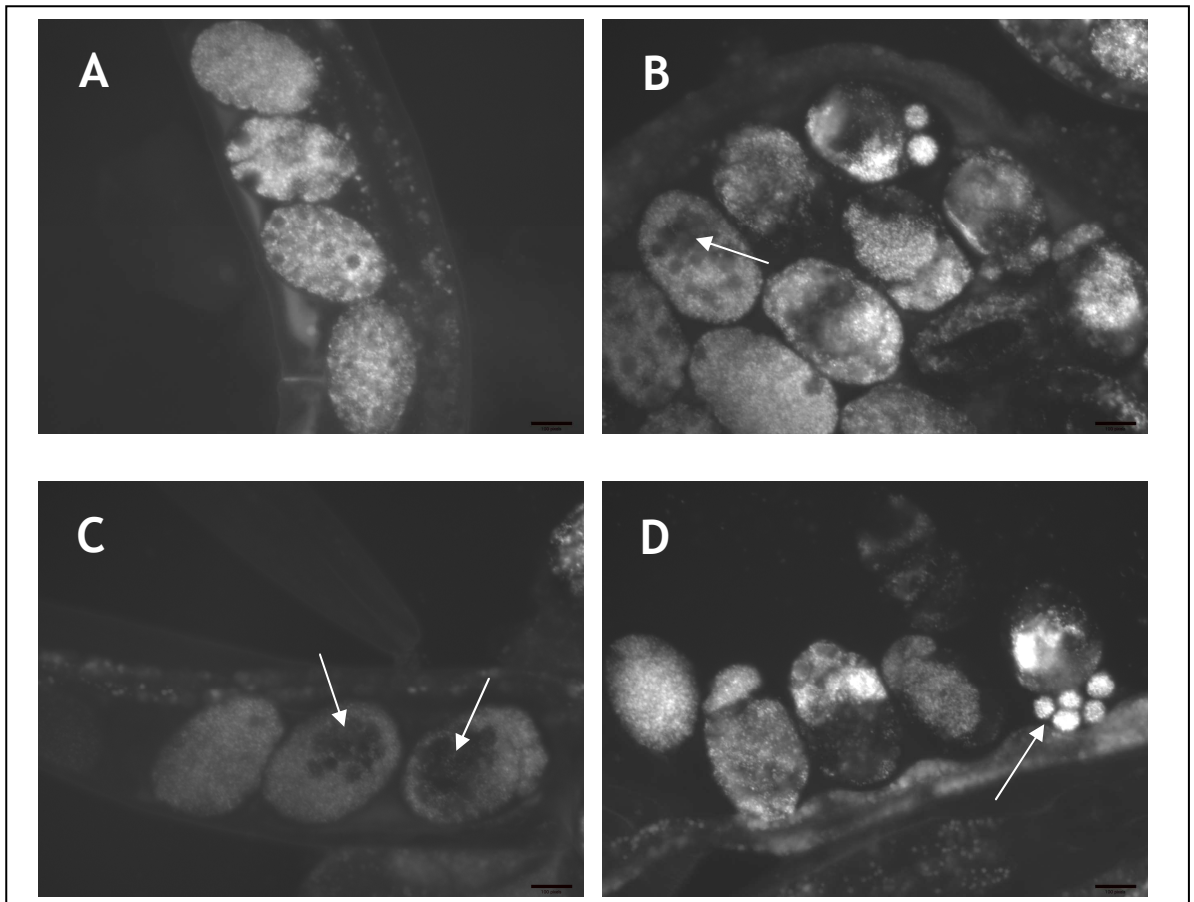
**Figure 6.11 GFP fluorescence in CLB01 embryos**

(A and B) CLB01 worms 3 days after the addition of L4 larvae to wells containing DMSO. Embryonic development is observed and also GFP fluorescence visible within the embryos. (C and D) CLB01 worms 3 days after the addition of L4 larvae to wells containing compound 1/20. Embryonic development is slower and GFP fluorescence minimal in embryos. Red circles indicate corresponding regions. All images at x40 magnification.



In an attempt to determine whether the compounds were effecting yolk production leading to embryonic effects, *C. elegans* strain BIS1 (*vit-2::gfp*) was examined. After exposure to the compounds, GFP fluorescence was no different from the control wells, suggesting that there is no direct effect on the production or uptake of yolk proteins. However, examination of the *vit-2::gfp* embryos indicated abnormal cell division, resulting in what appeared to be multinucleated cells and a number of abnormal small circular regions of cytoplasm (Figure 6.12).

It was suspected that the abnormal embryos would not develop properly, explaining the observed decrease in progeny. As the main aim of the drug screening work was to identify a compound that was having an effect on the ELT-2 GATA TF, an *elt-2* specific line (JR1838, *wls84pJM66 elt-2::gfp::lac-Z*, kindly supplied by Joel Rothman, University of California, Santa Barbara, CA) was also used to ensure that this was happening. This worm strain has strong GFP fluorescence in the gut nuclei and after exposure to compound, but not DMSO, in addition to the observed embryonic effects, there was a decrease in fluorescence suggesting that the compounds are also reducing *elt-2* (results not shown).



**Figure 6.12 *C. elegans* BIS1 (*vit-2::gfp*) worms indicating embryonic GFP fluorescence**

(A) BIS1 worms in control well, indicating normal division and distribution of yolk protein. (B, C and D) BIS1 worms exposed to compound 1/20 with arrows indicating pitted appearance and uneven yolk patterning. All at x40 magnification.

## 6.3 Discussion

The aim of this chapter was to identify compounds that potentially target the ELT-2 GATA transcription factor in *C. elegans* worms as indicated by a decrease in GFP expression from the *C. elegans cpl-1* promoter. Ideally compounds that inhibited the target would have a negative effect on growth and survival, with lethality being the eventual outcome. Work carried out by Couthier *et al.* (2004) identified an *elt-2* homologue in *H. contortus* capable of directing gut cell expression when ectopically expressed in *C. elegans*. It was speculated that compounds hitting the target may have a detrimental effect on *C. elegans* gut function and that the same may be observed in *H. contortus*.

The transgenic CLB01 worms exhibit GFP fluorescence in their gut and hypodermal cells, and *elt-2* RNAi results from this work (Chapter 5) suggest gut expression is under the control of the ELT-2 GATA TF. This ELT-2-directed GFP can be measured and used as a marker for ELT-2 activity, the key factor controlling gene transcription in the *C. elegans* intestine (McGhee *et al.*, 2007). Screening was carried out *in vitro*, testing fluorescence levels of transgenic CLB01 worms in M9 buffer, in the presence of a number of compounds. It was presumed that the level of fluorescence from the CLB01 worms would decrease during screening due to interference with gut gene expression and intestinal development in embryos and early larvae. It was anticipated that some decrease in fluorescence may occur as during maintenance of the CLB01 strain, GFP levels are reduced with extended culture and starvation. Compounds capable of interfering with ELT-2 activity may increase the speed of this effect. However, the level of fluorescence recorded in the first control readings taken by the EnVision did not show a significant decrease over the test period, but instead remained relatively constant over the 4 days. One explanation for the lack of decrease in fluorescence reading throughout the period of the screen was the presence of auto-fluorescence. Auto-fluorescence in *C. elegans* is a natural phenomenon and is associated with lipids, oxidative stress and related to the lifespan of the worm (Le *et al.*, 2010). As GFP and auto-fluorescence are observed at different locations in the *C. elegans* worm it is possible to distinguish between the two visually. However as they are very similar in their wavelength emissions it was difficult to use the EnVision to differentiate

between GFP and auto-fluorescence. Therefore it was imperative that all sample wells were observed microscopically in addition to the readings recorded by the EnVision. This was vitally important in those wells in which no measurable decrease in fluorescence was recorded by the EnVision, to ensure that any decrease in fluorescence was not being masked by the presence of auto-fluorescence.

Screens were carried out to determine the concentration of DMSO that could be tolerated by the transgenic CLB01 worm strain. In contrast to worms that became naturally starved or aged, in which auto-fluorescence was observed, no auto-fluorescence was observed in worms that had been killed by DMSO. Additionally GFP expression in worms killed by DMSO was at an equally high level at the start and end of the screen, i.e. worms still produced GFP fluorescence when dead. Thus it was assumed that because the worms had been killed rapidly by DMSO and had not become sick, the level of GFP fluorescence was not affected. Although a similar effect could not be assumed for other compounds, this observation suggested that where GFP levels decreased this may be a specific effect on gene expression and not a non-specific consequence of lethality.

The fluorescence that was observed from the OP50 bacteria presented an initial difficulty that had to be overcome. The fluorescence produced by bacteria has previously been identified as a problem when carrying out *C. elegans* screening, specifically if the results rely on a measure of fluorescence level (Gill *et al.*, 2003). The approach taken to combat this in the current study was to reduce the concentration of OP50 to a level that would not interfere with the fluorescence reading. This however often resulted in worms that were not as healthy as they would have been if grown on OP50 seeded plates or in OP50 culture rich media. As OP50 is the standard bacteria used for maintenance of *C. elegans* worms, finding a suitable alternative proved to be problematic. Work has been carried out in *C. elegans* using an anoxic medium, which is nutrient rich but contains no bacterial food source. When fed on this medium the worms had a slower rate of development and a reduced fecundity but an increase in lifespan was observed (Lenaerts *et al.*, 2008). This method of *C. elegans* feeding is not ideal as worms may not be at their most healthy, experiencing a degree of starvation, as suggested by the aforementioned phenotypes. However if the worms could be

maintained at a reasonable degree of health and for a longer period of time then this method of maintenance may be a solution to the interference fluorescence caused by OP50 bacteria when screening and could be tested for future screens.

As identified by the initial screen, the level of fluorescence recorded using the EnVision remained fairly constant for a period of up to 4 days. For this reason a control was required to confirm that if fluorescence in the worms decreased, this could be identified by the EnVision. There are no available compounds known to directly target the ELT-2 GATA TF to cause a decrease and for this reason *C. elegans elt-2* RNAi was used as an alternative. Feeding the worms on liquid bacteria expressing *C. elegans elt-2* dsRNA had not been tested before and it was possible that it would not be successful as worms do not feed as well when they are in the sample wells (comparable to liquid culture) as they do when maintained on plates (Muschiol *et al.*, 2009). Additionally, the low concentration of bacteria that was required so as to not interfere with fluorescence readings may not have provided the worms with enough bacteria to have a measurable RNAi effect on fluorescence.

RNAi results when worms were fed in the sample wells did not show a decrease in fluorescence, most likely for reasons discussed above. Therefore, no direct comparison between worms maintained in the sample wells on OP50 bacteria and on the *C. elegans elt-2* RNAi clone was possible. However a control was required to ensure that a decrease in fluorescence was measurable by the EnVision. The most commonly used RNAi method, feeding *C. elegans* worms on agar plates seeded with the RNAi clone, was therefore used. This method was successful and a decrease in fluorescence was observed.

One alternative that could have been attempted was soaking the *C. elegans* worms in dsRNA. The standard method involves soaking L4 larvae in dsRNA solution for a period of 24 hours, washing and then allowing them to feed on bacterial seeded plates to wait for the RNAi to have an effect (Maeda *et al.*, 2001). These conditions would not have been ideal for the type of fluorescence screening used here and modifications to this technique may have been more suited, for example soaking worms in dsRNA while in the sample wells would enable all stages of the worms to be targeted and for them to remain in the dsRNA for the entirety of the screen. This method however may have resulted in

decreased growth of young larvae as the ELT-2 GATA TF is essential for gut development and growth (Fukushige *et al.*, 1998). Although this would prove that RNAi was having the desired effect, it would not confirm that a decrease in fluorescence was measurable, as very young and unhealthy worms do not fluoresce. An alternative to combat this would be synchronisation of the worm population to silence *C. elegans elt-2* at the L4 stage.

The results of the *C. elegans* screen identified a number of compounds that could potentially be having an effect on the ELT-2 GATA TF target. These effects were determined by a decrease in the level of GFP fluorescence in the worms as recorded by the EnVision and also by a visual GFP decrease and worm death. All of the compounds that were screened in *C. elegans* were additionally screened in *H. contortus*. This enabled any compounds of interest from *C. elegans* studies to be examined in *H. contortus*. Ideally, based on functional conservation of the ELT-2 GATA TF in *H. contortus* (Couthier *et al.*, 2004), a similar effect may be observed in both nematodes. On analysis of these compounds in *H. contortus* none of them were found to be active. This was a disappointing finding, however each compound was only tested in one sample well and the recording of the results was subjective as it was carried out by eye. Repeating the screens with a higher number of sample wells may help identify if there is any activity that was missed in the initial screen. In addition compounds were tested only on the *H. contortus* L3 stage, which is not particularly active and may not take up a significant amount of compound. Other stages of *H. contortus* could therefore be examined. In the *C. elegans* screen, GFP was used as the predominant marker for identifying potential inhibition of the target and percent death was a second measure. 'Active' compounds in the *H. contortus* screen were those in which there was a high level of death only. For this reason the activity of compounds may be missed if they are inhibiting the target but having a slightly different effect e.g. starvation or reduced growth. If this is happening, these worms may not be easily identified when carrying out visual analysis of a large number of wells.

The method employed to confirm if an effect shown by a compound is real, rather than due to non-specific toxicity, is to generate a dose response curve for that compound, plotting drug concentration against percentage reduction in fluorescence and death, and looking for a correlation. It is reasonable to

presume that if a reduction is observed after 24 hours then this level of reduction would remain or be further reduced at the 48 hour time point. Of the compounds analysed in this way the graphs did not always follow this pattern and those which looked to be promising still had anomalies at different concentrations. Compound 2/10 was one example in which the graph did not follow the predicted pattern, with the fluorescent values at both time points remaining almost constant over the different concentrations. There are a number of reasons why the dose response curves did not always give the desired pattern. Only two wells were sampled at each concentration, therefore including more samples may have reduced the standard error. Additionally, at the lower concentrations even the smallest variation in the compound volume added to the well could have a dramatic effect on the results, due to the sensitivity of the assay. One compound, 2/7, did show a good correlation between drug concentration and effect.

There were three compounds that were identified as being active in *H. contortus* however none of these compounds were highlighted in the initial *C. elegans* screen. These compounds were re-screened in *C. elegans* for closer examination to determine if any effect may have been missed originally. Two of the compounds, X and Y, looked to cause a decrease in fluorescence in *C. elegans*. Statistical examination was carried out on the results, giving a 95% confidence in this decrease. Additionally, examination of the percentage death of *C. elegans* when exposed to these compounds showed a 50-60% death. In *H. contortus*, no effect was recorded at 24 hours and no observations are routinely made between this point and the final day 4 observation. For this reason it was unknown at exactly which point the compounds had produced an effect. *C. elegans* screens are relatively short in comparison to those carried out in *H. contortus* (2 days compared to 4 days). The main reason for this was that the screening conditions were not favourable for the maintenance of the *C. elegans* worms over an extended period. This problem could be addressed by using a different feeding media (as previously mentioned) in an attempt to maintain the health of the worms for a longer period of time. With longer exposure to the compound it may be possible to determine if a greater decrease in fluorescence and higher death rate could be observed, similar to that in *H. contortus*. Ideally, compounds that are inhibiting the ELT-2 GATA TF target are desired. Similar effects in both

species would allow full analysis of the mechanism of action using the *C. elegans* system.

*C. elegans* has traditionally been used as a model for *H. contortus* work due to ease of culture and genetic manipulation. For this reason it was presumed that using *C. elegans* for compound screening would be easier to work with than screening in *H. contortus*. However, establishing *C. elegans*-based screening presented several challenges. This type of screening required worms to be at optimum health to obtain maximum GFP expression. Worms require a constant food source (OP50 bacteria) and due to fluorescence from OP50, worms had to be extremely clean. This required ensuring that the worms and sample wells had minimal OP50 contamination, to ensure no interference with the results. This is in contrast to *H. contortus* larvae which can be easily prepared for screening and remain healthy for a number of days in the absence of bacteria. The ease with which *H. contortus* larvae can be maintained means that screening in *H. contortus* can be carried out for a much larger number of compounds as worm preparation is minimal and significantly higher numbers of worms are easily attainable. There are great differences in lifestyle between these two organisms. Inconsistencies in the reaction and potency to known anthelmintics has resulted in *C. elegans* not being widely used for whole parasite screening, however once a target of interest has been selected *C. elegans* is a very useful tool for mechanism-based screening to confirm its effects (Geary *et al.*, 1999). The advantage of using *C. elegans* worms over *H. contortus* is the ability to produce transgenic *C. elegans* worms. These worms can be used to confirm if compounds are selective for specific targets, whereas in *H. contortus* confirmation of mechanism of action is very difficult.

Use of *Ce-elt-2::gfp* strain JR1838 helped show compounds also reduced activity of ELT-2, supporting their action as possible TF inhibitors. In addition, more detailed analysis using *C. elegans* showed that compounds have a major effect on embryogenesis, with far fewer progeny being observed following exposure to compound.

In conclusion, the transgenic CLB01 worm strain proved to be a very useful tool in identification of compounds that potentially target the ELT-2 GATA TF. The developed screening method and conditions could easily be transferred and used



for the analysis of a number of GFP linked targets in the intestine or other cell types. Although compounds that appeared to have an effect on the desired target in *C. elegans* did not significantly affect *H. contortus* a possible reason for this may be lower drug uptake by *H. contortus*. Two compounds found to be effective at killing *H. contortus* were subsequently found to affect *C. elegans* *cpl-1* GFP expression and viability, therefore further examination of these in both species is warranted.

## **Chapter 7**

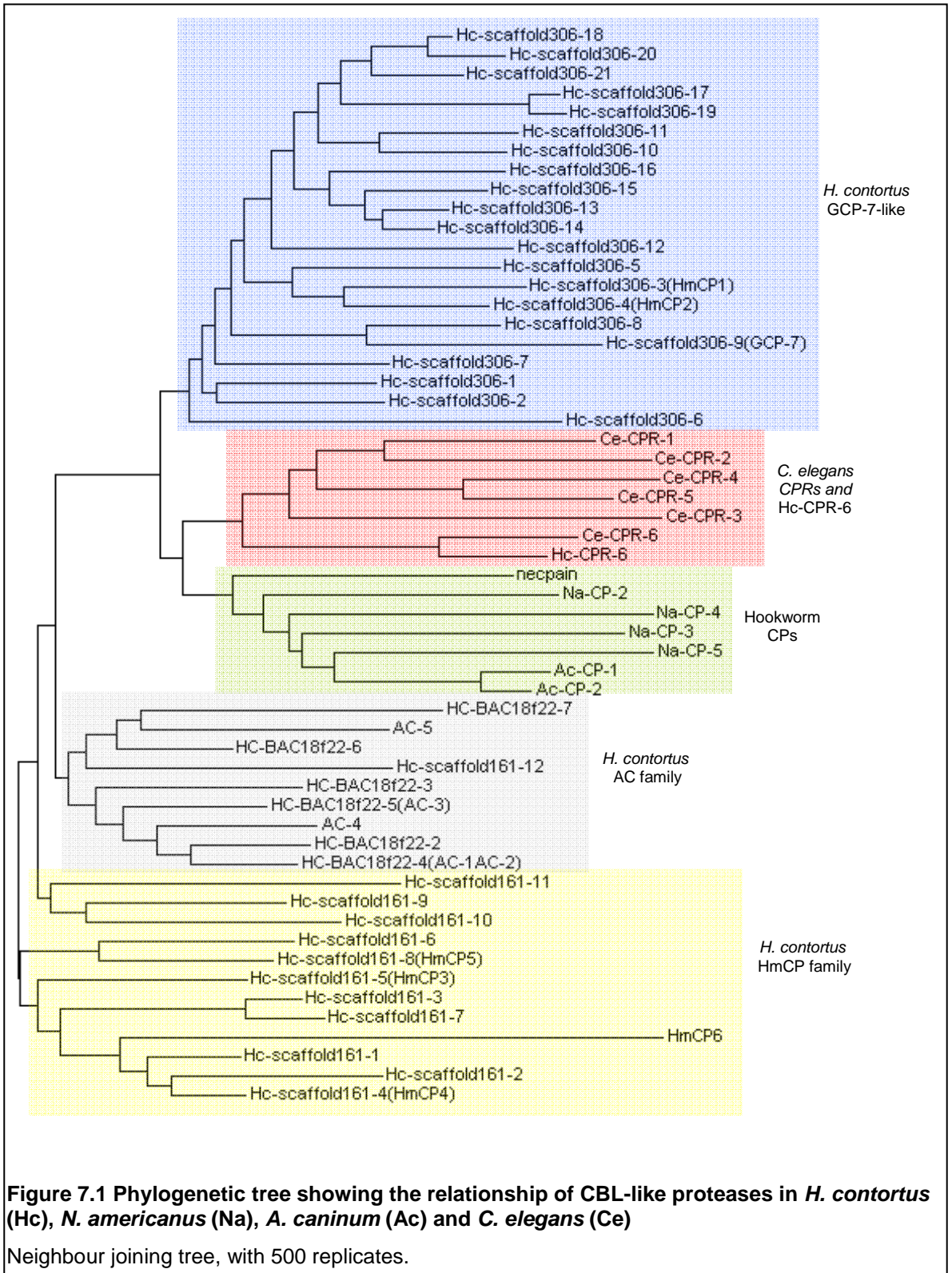
### General Discussion

Information about genes and genomes is continually increasing and with this comes invaluable advances in many fields. *H. contortus* is a worldwide economically important parasite, and due to the increasing resistance to anthelmintics much effort has been spent understanding the mechanisms of drug resistance and development of novel approaches to interfere with parasite development and survival (Knox *et al.*, 1999; Munn *et al.*, 1997). This is aided by advances in genome sequencing. The *H. contortus* genome sequencing project is currently ongoing and is anticipated to be completed in 2012/2013. The most recent assembly became available June 2012 and has an N50 of 83,238 bp (James Cotton, personal communication). Previous data had shown an N50 of 6,447 bp in May 2011 (Laing *et al.*, 2011), indicating the advances within this project and that the number and size of contiguous regions is rapidly increasing.

Multigene families have been found in a number of species. One study carried out by Demuth *et al.*, (2006) examined the evolution of mammalian gene families. When compared to the closely related chimpanzee genome, 6.4% of genes in the human genome were found to have no orthologs in the chimpanzee. This is also similar between the rat and mouse, with 10% of genes having no orthologs. A high rate of genetic turnover is accountable for these changes, with emphasis on gene duplication and deletion. It is therefore likely that gene families have undergone selective pressure and have altered to fill the requirements. Due to the ongoing sequencing of the *H. contortus* genome, there is no full comparative genomics with *C. elegans* or other nematodes. However, Yin *et al.*, (2008) carried out comparative analysis of intestinal transcriptomes of *A. suum*, *H. contortus* and *C. elegans*. There are over 5,000 known *C. elegans* intestinal genes which were compared to the 3,121 and 1,755 identified for *A. suum* and *H. contortus*, respectively. This was an extensive study that discovered 241 intestinal protein families having members within all three species. It is of interest to determine why some genes form expanded families and others are present in the genome as single copies. The expansion of CBL cysteine protease genes appears to occur more frequently than for other types of enzyme in *H. contortus*; for example the cathepsin L protease, Hc-CPL-1 (Britton and Murray, 2002), the metalloproteases MEP-1 and MEP-2 (Redmond *et al.*, 1997) and SOD (Liddell and Knox, 1998) are present as single copies. Additionally, Yin *et al.*, (Yin *et al.*, 2008) described 39% of *A. suum* and 19% of

*H. contortus* genes as being novel when compared to previously published databases, suggesting a significant degree of species-specific evolution.

In this thesis a number of multigene families in a number of species have been discussed. Ranjit *et al.*, (2008) described a family of cathepsin B cysteine proteases in the gut of the human hookworm *N. americanus* and in the dog hookworm *A. caninum*. Figure 7.1 is a phylogenetic tree indicating the relationship between all aforementioned proteases and families. From this figure it is evident that the cathepsin B-like gene family members in *H. contortus* are more similar to each other than to cathepsin Bs from other species, with the exception of Hc-CPR-6. Interestingly, CP proteases identified so far from hookworm species group more closely to *C. elegans* CPRs than to any of the *H. contortus* cathepsin B families. As hookworms also blood feed this suggests that the *H. contortus* CBLs have not diverged and duplicated specifically for blood-feeding, although further information on the full set of proteases in hookworms is needed to confirm this. In *H. contortus*, analysis to date has shown that the families are all located at different regions of the genome and it is likely that they have evolved to have different functions. One interesting point to note is the location of Hc-scaffold161-12 (HmCP genes) which appears to be more similar to the BAC 18f22 family than other genes on scaffold161.



A greater knowledge of members of gene families and their relationship to related families will help direct the development of novel anthelmintics or vaccines. In addition to CBLs, many studies have been carried out examining the importance of *H. contortus* H11 aminopeptidases, encoded by a family of tandemly arranged genes (Roberts and Johnston, unpublished data), the H-gal-GP complex and excretory-secretory (ES) components as potential vaccine

targets (Bakker *et al.*, 2004; Cachate *et al.*, 2010; De Vries *et al.*, 2009; Knox *et al.*, 1999; Munn *et al.*, 1997; Ruiz *et al.*, 2004a).

Cysteine protease genes are abundantly expressed in the *H. contortus* intestine and are thought to have a role in blood digestion, thus making them of interest as potential protective antigens. Work by Ruiz *et al.*, (2004b) examining five *H. contortus* cysteine proteases in two different strains identified 20 alleles. It is unknown what the implications of this genetic variation are, however changes may result in antigenic variation which may have an impact on the efficacy of potential novel vaccines. A recent study carried out by Molina *et al.*, (2012) reported a protective effect of cysteine protease-enriched fractions (CPFs) during a vaccine trial using two different *H. contortus* strains (North American and Spanish). This study reported a similar level of protection observed across the two strains, despite one strain being adapted to sheep and one to goats. It could be speculated that certain components within the CPFs with low genetic diversity e.g. AC-5 (Ruiz *et al.*, 2004b) may be maintaining a similar level of protection. It would be of interest to determine if there are specific cysteine protease genes that have low genetic diversity over a range of *H. contortus* strains, as these may be effective targets for a wide-scale novel vaccine. CPR-6 may be an obvious candidate; this protease is present only as a single copy and is highly conserved across species. No information is yet available on *cpr-6* sequences in different *H. contortus* strains, but this could be examined.

Bakker *et al.*, (2004) examined *H. contortus* protection using fractions from the adult stage excretory/secretory (ES) products. Analysis of the ES products identified a number of proteins that were un-characterised and it was thought that these may represent cysteine proteases. Interestingly, of the two ES fractions used in trials, the one containing the most cysteine protease activity gave the highest protection. Yatsuda *et al.*, (2006) also carried out a study examining the proteases in *H. contortus* ES products. Six different cysteine proteases were obtained; AC-4, GCP-7, Hm-CP1, Hm-CP1-like, Hm-CP2, and Hm-CP2-like. There were also three additional proteases identified, and as these did not show high DNA similarity to any known CBL were designated Hm-CP7 (CA869450, 181AA), Hm-CP8 (CB019057, ~53AA), Hm-CP9 (CA034108, 173AA). From the newly identified CBL sequences discussed in this thesis I have discovered that Hm-CP7 has 99% AA identity to Hc-scaffold161-1 and Hm-CP9

shows a 90% AA identity to Hc-scaffold306-10. Hm-CP8 only has a small amount of available sequence and therefore it is still unknown to which protein (if any) it is most similar. The work contained in this thesis has helped characterize and group many different cathepsin B proteases and supports the need for annotation and characterization of the complete *H. contortus* genome. This information would help in the identification of cysteine proteases within fractions such as the ES products (Bakker *et al.*, 2004) and subsequently help with protection studies, as targeting a number of proteins in a family may provide significant protection and be of benefit.

In addition to the work carried out characterising *H. contortus* gut cysteine proteases, in this thesis an assay was generated to test inhibition of a major gut TF, ELT-2. To date, screening potential anthelmintics in *H. contortus* is more frequently carried out than in the model organism *C. elegans*, as screening in the target organism is generally favoured. However, once a target of interest has been selected, *C. elegans* is a useful tool for mechanism-based screening to confirm compound effects (Geary *et al.*, 1999). *C. elegans* screening has recently been applied to human medicine, with screens being carried out for diseases such as Huntington's Disease (Voisine *et al.*, 2007) and Alzheimer's Disease (Lublin and Link, 2012). Gosai *et al.*, (2010) developed an assay examining the use of *C. elegans* in high-throughput screening. This work focussed on human liver disease caused by  $\alpha$ 1-antitrypsin deficiency and aimed to identify compounds that prevented misfolded protein accumulation, typical of the disease. The same difficulties identified in the generation of the screening assay in Chapter 6, such as DMSO tolerance and OP50 fluorescence, were also mentioned by Gosai *et al.*, (2010) but despite these factors an effective assay was produced.

This work provides an alternative to the generation of vaccines against gut proteases and protease families by targeting specific TFs involved in controlling their expression. TFs are essential for the control of specific genes and the roles of many nematode TFs, such as CEH-22 in pharyngeal muscles (Okkema and Fire, 1994) have been identified in *C. elegans* and the work of Reece-Hoys *et al.*, (2005), with 934 potential TFs identified in the *C. elegans* genome has provided much data for future work.

In conclusion, cathepsin B proteases in a number of parasitic species have been identified as having a place in the development of novel approaches to parasite control. In this thesis over 30 novel CBL genes have been identified and characterised. Two compounds were identified which can potentially target the ELT-2 TF in *C. elegans*, causing embryonic and larval effects. Further work testing these in *H. contortus* and other parasitic nematodes would therefore be of great interest.



## References

Abbott, K. A., Taylor, M., and Stubbings, L. A. (2009) *Sustainable Worm Control Strategies for Sheep 3rd Edition*.

Adcock, I. M. and Caramori, G. (2001) Cross-talk between pro-inflammatory transcription factors and glucocorticoids. *Immunology and Cell Biology* **79**: 376-384.

Allen, M. A., Hillier, L. W., Waterston, R. H., and Blumenthal, T. (2011) A global analysis of *C. elegans* trans-splicing. *Genome Research* **21**: 255-264.

Angulo-Cubillan, F. J., Garcia-Coiradas, L., Alunda, J. M., Cuquerella, M., and de la Fuente, C. (2010) Biological characterization and pathogenicity of three *Haemonchus contortus* isolates in primary infections in lambs. *Veterinary Parasitology* **171**: 99-105.

Bagga, S., Bracht, J., Hunter, S., Massirer, K., Holtz, J., Eachus, R. *et al.* (2005) Regulation by *let-7* and *lin-4* miRNAs results in target mRNA degradation. *Cell* **122**: 553-563.

Baig, S., Damian, R. T., and Peterson, D. S. (2006) A novel cathepsin B active site motif is shared by helminth bloodfeeders. *Experimental Parasitology* **101**: 83-89.

Bakker, N., Vervelde, L., Kanobana, K., Knox, D. P., Cornelissen, A. W. C. A., Vries, E. d. *et al.* (2004) Vaccination against the nematode *Haemonchus contortus* with a thiol-binding fraction from the excretory/secretory products (ES). *Vaccine* **22**: 619-629.

Banerjee, A. and Luduena, R. F. (1992) Kinetics of colchicine binding to purified beta-tubulin isotypes from bovine brain. *The Journal of Biological Chemistry* **267**: 13335-13339.

Barnes, P. J. and Adcock, I. M. (1998) Transcription factors and asthma. *European Respiratory Journal* **12**: 221-234.

Bartel, D. P. (2004) MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell* **116**: 281-297.

Bektesh, S., Van, D. K., and Hirsh, D. (1988) Presence of the *Caenorhabditis elegans* spliced leader on different mRNAs and in different genera of nematodes. *Genes and Development* **2**: 1277-1283.

Blaxter, M. (1998) *Caenorhabditis elegans* is a nematode. *Science* **282**: 2041-2046.

Bloom, L. and Horvitz, H. R. (1997) The *Caenorhabditis elegans* gene *unc-76* and its human homologs define a new gene family involved in axonal outgrowth and fasciculation. *Proceedings of the National Academy of Sciences of the United States of America, PNAS* **94**: 3414-3419.

- Blumenthal, T. (2005) Trans-splicing and operons. *WormBook* 1-9.
- Blumenthal, T. and Steward, K. (1997) RNA processing and gene structure. In C. *elegans* II. D.L.Riddle (ed.) Cold Spring Harbor: Cold Spring Harbor Laboratory Press: pp. 117-145.
- Bradnam, K. R. and Korf, I. (2008) Longer First Introns Are a General Property of Eukaryotic Gene Structure. *PLoS ONE* **3**: e3093.
- Brenner, S. (1974) The genetics of *Caenorhabditis elegans*. *Genetics* **77**: 71-94.
- Britton, C., McKerrow, J. H., and Johnstone, I. L. (1998) Regulation of the *Caenorhabditis elegans* gut cysteine protease gene *cpr-1*: requirement for GATA motifs. *Journal of Molecular Biology* **283**: 15-27.
- Britton, C. and Murray, L. (2002) A cathepsin L protease essential for *Caenorhabditis elegans* embryogenesis is functionally conserved in parasitic nematodes. *Molecular and Biochemical Parasitology* **122**: 21-33.
- Britton, C., Redmond, D. L., Knox, D. P., McKerrow, J. H., and Barry, J. D. (1999) Identification of promoter elements of parasite nematode genes in transgenic *Caenorhabditis elegans*. *Molecular Biochemical Parasitology* **103**: 171-181.
- Bruhat, A., Tourmente, S., Chapel, S., Sobrier, M. L., Couderc, J. L., and Dastugue, B. (1990) Regulatory elements in the first intron contribute to transcriptional regulation of the beta 3 tubulin gene by 20-hydroxyecdysone in *Drosophila* Kc cells. *Nucleic Acids Research* **18**: 2861-2867.
- Bürglin, T. R., Lobos, E., and Blaxter, M. L. (1998) *Caenorhabditis elegans* as a model for parasitic nematodes. *International Journal for Parasitology* **28**: 395-411.
- Burke, D. T., Carle, G. F., and Olson, M. V. (1987) Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236**: 806-812.
- Byerly, L., Cassada, R. C., and Russell, R. L. (1976) The life cycle of the nematode *Caenorhabditis elegans* : I. Wild-type growth and reproduction. *Developmental Biology* **51**: 23-33.
- Cachate, E., Newlands, G. F., Ekojsa, S. E., McAllister, H., and Smith, W. D. (2010) Attempts to immunize sheep against *Haemonchus contortus* using a cocktail of recombinant proteases derived from the protective antigen, H-gal-GP. *Parasite Immunology* **32**: 414-419.
- Callaghan, M. J. and Beh, K. J. (1994) Characterization of a tandemly repetitive DNA sequence from *Haemonchus contortus*. *International Journal for Parasitology* **24**: 137-141.
- Carthew, R. W. (2006) Gene regulation by microRNAs. *Current Opinion in Genetics & Development* **16**: 203-208.
- Coles, G. C. (2002) Sustainable use of anthelmintics in grazing animals. *Veterinary Record* **151**: 165-169.

- Conway, D. (1964) Variance in the effectiveness of Thiabendazole against *Haemonchus contortus* in sheep. *American Journal of Veterinary Research* **25**: 844-846.
- Coulson, A., Sulston, J., Brenner, S., and Karn, J. (1986) Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences* **83**: 7821-7825.
- Couthier, A., Smith, J., McGarr, P., Craig, B., and Gilleard, J. S. (2004) Ectopic expression of a *Haemonchus contortus* GATA transcription factor in *Caenorhabditis elegans* reveals conserved function in spite of extensive sequence divergence. *Molecular and Biochemical Parasitology* **133**: 241-253.
- Cox, G. N., Pratt, D., Hageman, R., and Boisvenue, R. J. (1990) Molecular cloning and primary sequence of a cysteine protease expressed by *Haemonchus contortus* adult worms. *Molecular and Biochemical Parasitology* **41**: 25-34.
- Cully, D. F., Wilkinson, H., Vassilatis, D. K., Etter, A., and Arena, J. P. (1996) Molecular biology and electrophysiology of glutamate-gated chloride channels of invertebrates. *Parasitology* **113 Suppl**: S191-S200.
- De Vries, E., Bakker, N., Krijgsveld, J., Knox, D. P., Heck, A. J., and Yatsuda, A. P. (2009) An AC-5 cathepsin B-like protease purified from *Haemonchus contortus* excretory secretory products shows protective antigen potential for lambs. *Veterinary Research* **40**: 41.
- Demuth, J., Bie, T., Stajich, J., Christianini, N., and Hahn, M. (2006) The Evolution of Mammalian Gene Families. *PLoS ONE* **1(1)**: e85.
- Deutsch, M. and Long, M. (1999) Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Research* **27**: 3219-3228.
- Devaney, E., Winter, A. D., and Britton, C. (2010) microRNAs: a role in drug resistance in parasitic nematodes? *Trends in Parasitology* **26**: 428-433.
- Drogemuller, M., Schnieder, T., and von Samson-Himmelstjerna, G. (2004) Beta-tubulin complementary DNA sequence variations observed between cyathostomins from benzimidazole-susceptible and -resistant populations. *Journal of Parasitology* **90**: 868-870.
- Ducray, P., Gauvry, N., Pautrat, F., Goebel, T., Fruechtel, J., Desaulles, Y. *et al.* (2008) Discovery of amino-acetonitrile derivatives, a new class of synthetic anthelmintic compounds. *Bioorganic & Medicinal Chemistry Letters* **18**: 2935-2938.
- Egan, C. R., Chung, M. A., Allen, F. L., Heschl, M. F., Van Buskirk, C. L., and McGhee, J. D. (1995) A gut-to-pharynx/tail switch in embryonic expression of the *Caenorhabditis elegans ges-1* gene centers on two GATA sequences. *Developmental Biology* **170**: 397-419.
- Emery, D. L. and Wagland, B. M. (1991) Vaccines against gastrointestinal nematode parasites of ruminants. *Parasitology Today* **7**: 347-349.

- Fire, A., Harrison, S. W., and Dixon, D. (1990) A modular set of lacZ fusion vectors for studying gene expression in *Caenorhabditis elegans*. *Gene* **93**: 189-198.
- Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998) Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**: 806-811.
- Fukushige, T., Hawkins, M. G., and McGhee, J. D. (1998) The GATA-factor *elt-2* is essential for formation of the *Caenorhabditis elegans* intestine. *Developmental Biology* **198**: 286-302.
- Fukushige, T., Goszczynski, B., Yan, J., and McGhee, J. D. (2005) Transcriptional control and patterning of the *pho-1* gene, an essential acid phosphatase expressed in the *C. elegans* intestine. *Developmental Biology* **279**: 446-461.
- Gao, X., Wang, Z., Martin, J., Abubucker, S., Zhang, X., Mitreva, M. *et al.* (2010) Identification of hookworm DAF-16/FOXO response elements and direct gene targets. *PLoS ONE* **19**;5: e12289.
- Gaudet, J. and McGhee, J. D. (2010) Recent advances in understanding the molecular mechanisms regulating *C. elegans* transcription. *Developmental Dynamics* **239**: 1388-1404.
- Geary, T. G., Conder, G. A., and Bishop, B. (2004) The changing landscape of antiparasitic drug discovery for veterinary medicine. *Trends in Parasitology* **20**: 449-455.
- Geary, T. G., Sangster, N. C., and Thompson, D. P. (1999) Frontiers in anthelmintic pharmacology. *Veterinary Parasitology* **84**: 275-295.
- Geldhof, P., Clark, D., Molloy, C., and Knox, D. P. (2007) Assessment of *Caenorhabditis elegans* as a model in *Haemonchus contortus* vaccine research. *Molecular Biochemical Parasitology* **152**: 220-223.
- Geldhof, P., Murray, L., Couthier, A., Gilleard, J. S., McLauchlan, G., Knox, D. P. *et al.* (2006) Testing the efficacy of RNA interference in *Haemonchus contortus*. *International Journal for Parasitology* **36**: 801-810.
- Geldhof, P., Whitton, C., Gregory, W. F., Blaxter, M., and Knox, D. P. (2005) Characterisation of the two most abundant genes in the *Haemonchus contortus* expressed sequence tag dataset. *International Journal for Parasitology* **35**: 513-522.
- Ghedini, E., Wang, S., Spiro, D., Caler, E., Zhao, Q., Crabtree, J. *et al.* (2007) Draft genome of the filarial nematode parasite *Brugia malayi*. *Science* **317**: 1756-1760.
- Gill, M. S., Olsen, A., Sampayo, J. N., and Lithgow, G. J. (2003) An automated high-throughput assay for survival of the nematode *Caenorhabditis elegans*. *Free Radical Biology and Medicine* **35**: 558-565.

- Gillan, V., Maitland, K., McCormack, G., Him, N. A., and Devaney, E. (2009) Functional genomics of *hsp-90* in parasitic and free-living nematodes. *International Journal for Parasitology* **39**: 1071-1081.
- Gilleard, J. S. (2004) The use of *Caenorhabditis elegans* in parasitic nematode research. *Parasitology* **128 Suppl 1**: S49-S70.
- Gilleard, J. S., Woods, D. J., and Dow, J. A. T. (2005) Model-organism genomics in veterinary parasite drug-discovery. *Trends in Parasitology* **21**: 302-305.
- Githigia, S. M., Thamsborg, S. M., Munyua, W. K., and Maingi, N. (2001) Impact of gastrointestinal helminths on production in goats in Kenya. *Small Ruminant Research* **42**: 21-29.
- Gomez-Escobar, N., Gregory, W. F., Britton, C., Murray, L., Corton, C., Hall, N. *et al.* (2002) Abundant larval *transcript-1* and *-2* genes from *Brugia malayi*: diversity of genomic environments but conservation of 5' promoter sequences functional in *Caenorhabditis elegans*. *Molecular and Biochemical Parasitology* **125**: 59-71.
- Gosai, S. J., Kwak, J. H., Luke, C. J., Long, O. S., King, D. E., Kovatch, K. J. *et al.* (2010) Automated high-content live animal drug screening using *C. elegans* expressing the aggregation prone serpin *alpha1-antitrypsin Z*. *PLoS ONE* **5**: e15460.
- Guenette, R. S., Mooibroek, M., Wong, K., Wong, P., and Tenniswood, M. (1994) Cathepsin B, a cysteine protease implicated in metastatic progression, is also expressed during regression of the rat prostate and mammary glands. *European Journal of Biochemistry* **226**: 311-321.
- Haerty, W., Artieri, C., Khezri, N., Singh, R. S., and Gupta, B. P. (2008) Comparative analysis of function and interaction of transcription factors in nematodes: extensive conservation of orthology coupled to rapid sequence evolution. *BMC Genomics* **9**: 399.
- Harrop, S. A., Sawangjaroen, N., Prociw, P., and Brindley, P. J. (1995) Characterization and localization of cathepsin B proteinases expressed by adult *Ancylostoma caninum* hookworms. *Molecular and Biochemical Parasitology* **71**: 163-171.
- Harrow, I. D. and Gratton, K. A. F. (1985) Mode of action of the anthelmintics morantel, pyrantel and levamisole on muscle cell membrane of the nematode *Ascaris suum*. *Pesticide Science* **16**: 662-672.
- Hashmi, S., Britton, C., Liu, J., Guiliano, D. B., Oksov, Y., and Lustigman, S. (2002) Cathepsin L is essential for embryogenesis and development of *Caenorhabditis elegans*. *The Journal of Biological Chemistry* **277**: 3477-3486.
- Hawkins, M. G. and McGhee, J. D. (1995) *elt-2*, a second GATA factor from the nematode *Caenorhabditis elegans*. *The Journal of Biological Chemistry* **270**: 14666-14671.

- Heussler, V. T. and Dobbela, D. A. E. (1994) Cloning of a protease gene family of *Fasciola hepatica* by the polymerase chain reaction. *Molecular and Biochemical Parasitology* **64**: 11-23.
- Ho, S. H., So, G. M., and Chow, K. L. (2001) Postembryonic expression of *Caenorhabditis elegans* *mab-21* and its requirement in sensory ray differentiation. *Developmental Dynamics* **221**: 422-430.
- Hobert, O. (2008) Gene regulation by transcription factors and microRNAs. *Science* **319**: 1785-1786.
- Hoekstra, R., Criado-Fornelio, A., Fakkeldij, J., Bergman, J., and Roos, M. H. (1997) Microsatellites of the parasitic nematode *Haemonchus contortus*: polymorphism and linkage with a direct repeat. *Molecular and Biochemical Parasitology* **89**: 97-107.
- Hotez, P. J., Brindley, P. J., Bethony, J. M., King, C. H., Pearce, E. J., and Jacobson, J. (2008) Helminth infections: the great neglected tropical diseases. *The Journal of Clinical Investigation* **118**: 1311-1321.
- Hsu, A. L., Murphy, C. T., and Kenyon, C. (2003) Regulation of Aging and Age-Related Disease by DAF-16 and Heat-Shock Factor. *Science* **300**: 1142-1145.
- Hughes, A. L. (1994) The evolution of functionally novel proteins after gene duplication. *Proceedings of the Royal Society Biological Sciences* **256**: 119-124.
- Illy, C., Quraishi, O., Wang, J., Purisima, E., Vernet, T., and Mort, J. S. (1997) Role of the occluding loop in cathepsin B activity. *The Journal of Biological Chemistry* **272**: 1197-1202.
- Jasmer, D. P. and McGuire, T. C. (1991) Protective immunity to a blood-feeding nematode (*Haemonchus contortus*) induced by parasite gut antigens. *Infection and Immunity* **59**: 4412-4417.
- Jasmer, D. P., Mitreva, M. D., and McCarter, J. P. (2004) mRNA sequences for *Haemonchus contortus* intestinal cathepsin B-like cysteine proteases display an extreme in abundance and diversity compared with other adult mammalian parasitic nematodes. *Molecular and Biochemical Parasitology* **137**: 297-305.
- Jasmer, D. P., Roth, J., and Myler, P. J. (2001) Cathepsin B-like cysteine proteases and *Caenorhabditis elegans* homologues dominate gene products expressed in adult *Haemonchus contortus* intestine. *Molecular and Biochemical Parasitology* **116**: 159-169.
- Jeong, Y. M., Mun, J. H., Lee, I., Woo, J. C., Hong, C. B., and Kim, S. G. (2006) Distinct roles of the first introns on the expression of *Arabidopsis* profilin gene family members. *Plant Physiology* **140**: 196-209.
- Jordan, V. C. (2003) Tamoxifen: a most unlikely pioneering medicine. *Nature Reviews Drug Discovery* **2**: 205-213.
- Juo, Z. S., Chiu, T. K., Leiberman, P. M., Baikalov, I., Berk, A. J., and Dickerson, R. E. (1996) How proteins recognize the TATA box. *Journal of Molecular Biology* **261**: 239-254.

- Kamath, R. S. and Ahringer, J. (2003) Genome-wide RNAi screening in *Caenorhabditis elegans*. *Methods* **30**: 313-321.
- Kaminsky, R., Ducray, P., Jung, M., Clover, R., Rufener, L., Bouvier, J. *et al.* (2008a) A new class of anthelmintics effective against drug-resistant nematodes. *Nature* **452**: 176-180.
- Kaminsky, R., Gauvry, N., Schorderet Weber, S., Skripsky, T., Bouvier, J., Wenger, A. *et al.* (2008b) Identification of the amino-acetonitrile derivative monepantel (AAD 1566) as a new anthelmintic drug development candidate. *Parasitology Research* **103**: 931-949.
- Kaminsky, R., Mosimann, D., Sager, H., Stein, P., and Hosking, B. (2009) Determination of the effective dose rate for monepantel (AAD 1566) against adult gastro-intestinal nematodes in sheep. *International Journal for Parasitology* **39**: 443-446.
- Kaplan, R. M. (2004) Drug resistance in nematodes of veterinary importance: a status report. *Trends in Parasitology* **20**: 477-481.
- Kennedy, B. P., Aamodt, E. J., Allen, F. L., Chung, M. A., Heschl, M. F., and McGhee, J. D. (1993) The gut esterase gene (*ges-1*) from the nematodes *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Journal of Molecular Biology* **20**;229: 890-908.
- Kenyon, F., Sargison, N. D., Skuce, P. J., and Jackson, F. (2009) Sheep helminth parasitic disease in south eastern Scotland arising as a possible consequence of climate change. *Veterinary Parasitology* **163**: 293-297.
- Klinkert, M. Q., Felleisen, R., Link, G., Ruppel, A., and Beck, E. (1989) Primary structures of Sm31/32 diagnostic proteins of *Schistosoma mansoni* and their identification as proteases. *Molecular and Biochemical Parasitology* **33**: 113-122.
- Knox, D. P., Redmond, D. L., Newlands, G. F., Skuce, P. J., Pettit, D., and Smith, W. D. (2003) The nature and prospects for gut membrane proteins as vaccine candidates for *Haemonchus contortus* and other ruminant trichostrongyloids. *International Journal for Parasitology* **33**: 1129-1137.
- Knox, D. P., Smith, S. K., and Smith, W. D. (1999) Immunization with an affinity purified protein extract from the adult parasite protects lambs against infection with *Haemonchus contortus*. *Parasite Immunology* **21**: 201-210.
- Kramer, J. M., French, R. P., Park, E. C., and Johnson, J. J. (1990) The *Caenorhabditis elegans* *rol-6* gene, which interacts with the *sqt-1* collagen gene to determine organismal morphology, encodes a collagen. *Molecular and Cellular Biology* **10**: 2081-2089.
- Kwa, M. S., Veenstra, J. G., Van Dijk, M., and Roos, M. H. (1995) Beta-tubulin genes from the parasitic nematode *Haemonchus contortus* modulate drug resistance in *Caenorhabditis elegans*. *Journal of Molecular Biology* **246**: 500-510.



- Lacey, E. (1988) The role of the cytoskeletal protein, tubulin, in the mode of action and mechanism of drug resistance to benzimidazoles. *International Journal for Parasitology* **18**: 885-936.
- Lacey, E. (1990) Mode of action of benzimidazoles. *Parasitology Today* **6**: 112-115.
- Laing, R., Hunt, M., Protasio, A. V., Saunders, G., Mungall, K., Laing, S. *et al.* (2011) Annotation of Two Large Contiguous Regions from the *Haemonchus contortus* Genome Using RNA-seq and Comparative Analysis with *Caenorhabditis elegans*. *PLoS ONE* **6**: e23216.
- Lall, S., Friedman, C. C., Jankowska-Anyszka, M., Stepinski, J., Darzynkiewicz, E., and Davis, R. E. (2004) Contribution of trans-splicing, 5' -leader length, cap-poly(A) synergism, and initiation factors to nematode translation in an *Ascaris suum* embryo cell-free system. *Journal of Biological Chemistry* **279**: 45573-45585.
- Lanusse, C. E. and Prichard, R. K. (1993) Relationship between pharmacological properties and clinical efficacy of ruminant anthelmintics. *Veterinary Parasitology* **49**: 123-158.
- Larminie, C. G. and Johnstone, I. L. (1996) Isolation and characterization of four developmentally regulated cathepsin B-like cysteine protease genes from the nematode *Caenorhabditis elegans*. *DNA Cell Biology* **15**: 75-82.
- Latchman, D. S. (2000) Transcription factors as potential targets for therapeutic drugs. *Current Pharmaceutical Biotechnology* **1**: 57-61.
- Le, T. T., Duren, H. M., Slipchenko, M. N., Hu, C. D., and Cheng, J. X. (2010) Label-free quantitative analysis of lipid metabolism in living *Caenorhabditis elegans*. *The Journal of Lipid Research* **51**: 672-677.
- Lee, B. H., Clothier, M. F., Dutton, F. E., Nelson, S. J., Johnson, S. S., Thompson, D. P. *et al.* (2002) Marcfortine and paraherquamide class of anthelmintics: discovery of PNU-141962. *Current Topics in Medicinal Chemistry* **2**: 779-793.
- Lee, R. C., Feinbaum, R. L., and Ambros, V. (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**: 843-854.
- Lenaerts, I., Walker, G. A., Van Hoorebeke, L., Gems, D., and Vanfleteren, J. R. (2008) Dietary Restriction of *Caenorhabditis elegans* by Axenic Culture Reflects Nutritional Requirement for Constituents Provided by Metabolically Active Microbes. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **63**: 242-252.
- Leroy, S., Duperray, C., and Morand, S. (2003) Flow cytometry for parasite nematode genome size measurement. *Molecular and Biochemical Parasitology* **128**: 91-93.

- Lewis, J. A., Wu, C. H., Levine, J. H., and Berg, H. (1980) Levamisole-resistant mutants of the nematode *Caenorhabditis elegans* appear to lack pharmacological acetylcholine receptors. *Neuroscience* **5**: 967-989.
- Li, X., Massey, H. C., Jr., Nolan, T. J., Schad, G. A., Kraus, K., Sundaram, M. *et al.* (2006) Successful transgenesis of the parasitic nematode *Strongyloides stercoralis* requires endogenous non-coding control elements. *International Journal for Parasitology* **36**: 671-679.
- Libina, N., Berman, J. R., and Kenyon, C. (2003) Tissue-specific activities of *C. elegans* DAF-16 in the regulation of lifespan. *Cell* **115**: 489-502.
- Liddell, S. and Knox, D. P. (1998) Extracellular and cytoplasmic Cu/Zn superoxide dismutases from *Haemonchus contortus*. *Parasitology* **116**: 383-394.
- Little, P. R., Hodge, A., Maeder, S. J., Wirtherle, N. C., Nicholas, D. R., Cox, G. G. *et al.* (2011) Efficacy of a combined oral formulation of derquantel-abamectin against the adult and larval stages of nematodes in sheep, including anthelmintic-resistant strains. *Veterinary Parasitology* **181**: 180-193.
- Little, P. R., Hodges, A., Watson, T. G., Seed, J. A., and Maeder, S. J. (2010) Field efficacy and safety of an oral formulation of the novel combination anthelmintic, derquantel-abamectin, in sheep in New Zealand. *New Zealand Veterinary Journal* **58**: 121-129.
- Liu, L., Panangala, V. S., and Dybvig, K. (2002) Trinucleotide GAA Repeats Dictate pMGA Gene Expression in *Mycoplasma gallisepticum* by Affecting Spacing between Flanking Regions. *Journal of Bacteriology*.
- Loukas, A., Bethony, J. M., Williamson, A. L., Goud, G. N., Mendez, S., Zhan, B. *et al.* (2004) Vaccination of dogs with a recombinant cysteine protease from the intestine of canine hookworms diminishes the fecundity and growth of worms. *The Journal of Infectious Diseases* **189**: 1952-1961.
- Lublin, A. L. and Link, C. D. (2012) Alzheimer's disease drug discovery: in vivo screening using *Caenorhabditis elegans* as a model for  $\beta$ -amyloid peptide-induced toxicity. *Drug Discovery Today: Technologies*.
- MacMorris, M., Broverman, S., Greenspoon, S., Lea, K., Madej, C., Blumenthal, T. *et al.* (1992) Regulation of vitellogenin gene expression in transgenic *Caenorhabditis elegans*: short sequences required for activation of the *vit-2* promoter. *Molecular and Cellular Biology* **12**: 1652-1662.
- Maeda, I., Kohara, Y., Yamamoto, M., and Sugimoto, A. (2001) Large-scale analysis of gene function in *Caenorhabditis elegans* by high-throughput RNAi. *Current Biology* **11**: 171-176.
- Martinez, N. J., Ow, M. C., Barrasa, M. I., Hammell, M., Sequerra, R., Doucette-Stamm, L. *et al.* (2008) A *C. elegans* genome-scale microRNA network contains composite feedback motifs with high flux capacity. *Genes and Development* **22**: 2535-2549.

- Mascarenhas, D., Mettler, I. J., Pierce, D. A., and Lowe, H. W. (1990) Intron-mediated enhancement of heterologous gene expression in maize. *Plant Molecular Biology* **15**: 913-920.
- McCracken, R. O. and Lipkowitz, K. B. (1990) Structure-activity relationships of benzothiazole and benzimidazole anthelmintics: a molecular modeling approach to *in vivo* drug efficacy. *Journal of Parasitology* **76**: 853-864.
- McGhee, J. D., Sleumer, M. C., Bilenky, M., Wong, K., McKay, S. J., Goszczynski, B. *et al.* (2007) The ELT-2 GATA-factor and the global regulation of transcription in the *C. elegans* intestine. *Developmental Biology* **302**: 627-645.
- Melendez, A., Tallochy, Z., Seaman, M., Eskelinen, E. L., Hall, D. H., and Levine, B. (2003) Autophagy genes are essential for dauer development and life-span extension in *C. elegans*. *Science* **301**: 1387-1391.
- Miller, J. R., Koren, S., and Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. *Genomics* **95**: 315-327.
- Molina, J. M., Martin, S., Hernandez, Y. I., Gonzalez, J. F., Ferrer, O., and Ruiz, A. (2012) Immunoprotective effect of cysteine proteinase fractions from two *Haemonchus contortus* strains adapted to sheep and goats. *Veterinary Parasitology* **188**: 53-59.
- Mottram, J. C., Brooks, D. R., and Coombs, G. H. (1998) Roles of cysteine proteinases of trypanosomes and *Leishmania* in host-parasite interactions. *Current Opinion in Microbiology* **1**: 455-460.
- Munn, E. A., Greenwood, C. A., and Coadwell, W. J. (1987) Vaccination of young lambs by means of a protein fraction extracted from adult *Haemonchus contortus*. *Parasitology* **94** ( Pt 2): 385-397.
- Munn, E. A., Smith, T. S., Smith, H., James, F. M., Smith, F. C., and Andrews, S. J. (1997) Vaccination against *Haemonchus contortus* with denatured forms of the protective antigen H11. *Parasite Immunology* **19**: 243-248.
- Murray, L., Geldhof, P., Clark, D., Knox, D. P., and Britton, C. (2007) Expression and purification of an active cysteine protease of *Haemonchus contortus* using *Caenorhabditis elegans*. *International Journal for Parasitology* **37**: 1117-1125.
- Muschiol, D., Schroeder, F., and Traunspurger, W. (2009) Life cycle and population growth rate of *Caenorhabditis elegans* studied by a new method. *BMC Ecology* **9**:14.: 14.
- Nagaraj, S. H., Gasser, R. B., and Ranganathan, S. (2007) A hitchhiker's guide to expressed sequence tag (EST) analysis. *Briefings in Bioinformatics* **8**: 6-21.
- Nam, S., Jin, Y. H., Li, Q. L., Lee, K. Y., Jeong, G. B., Ito, Y. *et al.* (2002) Expression Pattern, Regulation, and Biological Role of Runt Domain Transcription Factor, *run*, in *Caenorhabditis elegans*. *Molecular and Cellular Biology* **22**: 547-554.

- O'Shea-Greenfield, A. and Smale, S. T. (1992) Roles of TATA and initiator elements in determining the start site location and direction of RNA polymerase II transcription. *Journal of Biological Chemistry* **267**: 1391-1402.
- Okkema, P. G. and Fire, A. (1994) The *Caenorhabditis elegans* NK-2 class homeoprotein CEH-22 is involved in combinatorial activation of gene expression in pharyngeal muscle. *Development* **120**: 2175-2186.
- Okkema, P. G., Harrison, S. W., Plunger, V., Aryana, A., and Fire, A. (1993) Sequence requirements for myosin gene expression and regulation in *Caenorhabditis elegans*. *Genetics* **135**: 385-404.
- Ongvarrasoponea, C., Roshorma, Y., and Panyima, A. (2007) A Simple and Cost Effective Method to Generate dsRNA for RNAi Studies in Invertebrates. *ScienceAsia* **33**: 35-39.
- Oxberry, M. E., Geary, T. G., and Prichard, R. K. (2001) Assessment of benzimidazole binding to individual recombinant tubulin isotypes from *Haemonchus contortus*. *Parasitology* **122**: 683-687.
- Parkinson, J., Guiliano, D. B., and Blaxter, M. (2002) Making sense of EST sequences by CLOBBing them. *BMC Bioinformatics* **3**:31.: 31.
- Pillai, S., Kalinna, B. H., Liebau, E., Hartmann, S., Theuring, F., and Lucius, R. (2005) Studies on *Acanthocheilonema viteae* cystatin: Genomic organization, promoter studies and expression in *Caenorhabditis elegans*. *Filaria Journal* **4**: 9.
- Pratt, D., Armes, L. G., Hageman, R., Reynolds, V., Boisvenue, R. J., and Cox, G. N. (1992) Cloning and sequence comparisons of four distinct cysteine proteases expressed by *Haemonchus contortus* adult worms. *Molecular and Biochemical Parasitology* **51**: 209-218.
- Pratt, D., Cox, G. N., Milhausen, M. J., and Boisvenue, R. J. (1990) A developmentally regulated cysteine protease gene family in *Haemonchus contortus*. *Molecular Biochemical Parasitology* **43**: 181-191.
- Prichard, R. (2001) Genetic variability following selection of *Haemonchus contortus* with anthelmintics. *Trends in Parasitology* **17**: 445-453.
- Prichard, R. K. (1990) Anthelmintic resistance in nematodes: Extent, recent understanding and future directions for control and research. *International Journal for Parasitology* **20**: 515-523.
- Ranjit, N., Zhan, B., Stenzel, D. J., Mulvenna, J., Fujiwara, R., Hotez, P. J. *et al.* (2008) A family of cathepsin B cysteine proteases expressed in the gut of the human hookworm, *Necator americanus*. *Molecular Biochemical Parasitology* **160**: 90-99.
- Ray, C. and McKerrow, J. H. (1992) Gut-specific and developmental expression of a *Caenorhabditis elegans* cysteine protease gene. *Molecular and Biochemical Parasitology* **51**: 239-249.

Redmond, D. L. and Knox, D. P. (2004) Protection studies in sheep using affinity-purified and recombinant cysteine proteinases of adult *Haemonchus contortus*. *Vaccine* **22**: 4252-4261.

Redmond, D. L. and Knox, D. P. (2006) Further protection studies using recombinant forms of *Haemonchus contortus* cysteine proteinases. *Parasite Immunology* **28**: 213-219.

Redmond, D. L., Knox, D. P., Newlands, G., and Smith, W. D. (1997) Molecular cloning and characterisation of a developmentally regulated putative metallopeptidase present in a host protective extract of *Haemonchus contortus*. *Molecular Biochemical Parasitology* **85**: 77-87.

Reece-Hoyes, J. S., Deplancke, B., Shingles, J., Grove, C. A., Hope, I. A., and Walhout, A. J. (2005) A compendium of *Caenorhabditis elegans* regulatory transcription factors: a resource for mapping transcription regulatory networks. *Genome Biology* **6**: R110.

Rehman, A. and Jasmer, D. P. (1998) A tissue specific approach for analysis of membrane and secreted protein antigens from *Haemonchus contortus* gut and its application to diverse nematode species. *Molecular Biochemical Parasitology* **97**: 55-68.

Rehman, A. and Jasmer, D. P. (1999) Defined characteristics of cathepsin B-like proteins from nematodes: inferred functional diversity and phylogenetic relationships. *Molecular Biochemical Parasitology* **102**: 297-310.

Rispe, C., Kutsukake, M., Doublet, V., Hudaverdian, S., Legeai, F., Simon, J. C. *et al.* (2008) Large gene family expansion and variable selective pressures for cathepsin B in aphids. *Molecular Biology and Evolution* **25**: 5-17.

Rose, A. B., Elfersi, T., Parra, G., and Korf, I. (2008) Promoter-proximal introns in *Arabidopsis thaliana* are enriched in dispersed signals that elevate gene expression. *The Plant Cell* **20**: 543-551.

Rufener, L., Keiser, J., Kaminsky, R., Maser, P., and Nilsson, D. (2010) Phylogenomics of ligand-gated ion channels predicts monepantel effect. *PLoS Pathogens* **6**: e1001091.

Ruiz, A., Molina, J. M., Gonzalez, J. F., Conde, M. M., Martin, S., and Hernandez, Y. I. (2004a) Immunoprotection in goats against *Haemonchus contortus* after immunization with cysteine protease enriched protein fractions. *Veterinary Research* **35**: 565-572.

Ruiz, A., Molina, J. M., Njue, A., and Prichard, R. K. (2004b) Genetic variability in cysteine protease genes of *Haemonchus contortus*. *Parasitology* **128**: 549-559.

Schallig, H. D., van Leeuwen, M. A., and Cornelissen, A. W. (1997) Protective immunity induced by vaccination with two *Haemonchus contortus* excretory secretory proteins in sheep. *Parasite Immunology* **19**: 447-453.

Sharma, R. L., Bhat, T. K., and Dhar, D. N. (1988) Control of sheep lungworm in India. *Parasitology Today* **4**: 33-36.

- Shi, W. and Zhou, W. (2006) Frequency distribution of TATA Box and extension sequences on human promoters. *BMC Bioinformatics* **7 Suppl 4:S2**: S2.
- Shompole, S. and Jasmer, D. P. (2001) Cathepsin B-like cysteine proteases confer intestinal cysteine protease activity in *Haemonchus contortus*. *Journal of Biological Chemistry* **276**: 2928-2934.
- Skuce, P. J., Redmond, D. L., Liddell, S., Stewart, E. M., Newlands, G. F., Smith, W. D. *et al.* (1999) Molecular cloning and characterization of gut-derived cysteine proteinases associated with a host protective extract from *Haemonchus contortus*. *Parasitology* **119 ( Pt 4)**: 405-412.
- Sleigh, J. N. (2010) Functional analysis of nematode nicotinic receptors. *Bioscience Horizons* **3**: 29-39.
- Smith, T. S., Graham, M., Munn, E. A., Newton, S. E., Knox, D. P., Coadwell, W. J. *et al.* (1997) Cloning and characterization of a microsomal aminopeptidase from the intestine of the nematode *Haemonchus contortus*. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* **1338**: 295-306.
- Smith, W. D., Skuce, P. J., Newlands, G. F., Smith, S. K., and Pettit, D. (2003) Aspartyl proteases from the intestinal brush border of *Haemonchus contortus* as protective antigens for sheep. *Parasite Immunology* **25**: 521-530.
- Smith, W. D., Smith, S. K., Pettit, D., Newlands, G. F., and Skuce, P. J. (2000) Relative protective properties of three membrane glycoprotein fractions from *Haemonchus contortus*. *Parasite Immunology* **22**: 63-71.
- Spieth, J., Shim, Y. H., Lea, K., Conrad, R., and Blumenthal, T. (1991) *elt-1*, an embryonically expressed *Caenorhabditis elegans* gene homologous to the GATA transcription factor family. *Molecular and Cellular Biology* **11**: 4651-4659.
- Stainier, D. Y. (2002) A glimpse into the molecular entrails of endoderm formation. *Genes and Development* **16**: 893-907.
- Stein, L. D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M. R., Chen, N. *et al.* (2003) The genome sequence of *Caenorhabditis briggsae*: a platform for comparative genomics. *PLoS Biology* **1**: E45.
- Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R. *et al.* (1992) The *C. elegans* genome sequencing project: a beginning. *Nature* **356**: 37-41.
- Tabara, H., Grishok, A., and Mello, C. C. (1998) RNAi in *C. elegans*: Soaking in the Genome Sequence. *Science* **282**: 430-431.
- Taylor, M. (1999) Use of anthelmintics in sheep. *In Practice* **21**: 222-231.
- Taylor, M. (2009) Changing patterns of parasitism in sheep. *In Practice* **31**: 474-483.
- Thamsborg, S. M., Jørgensen, R. J., Waller, P. J., and Nansen, P. (1996) The influence of stocking rate on gastrointestinal nematode infections of sheep over a 2-year grazing period. *Veterinary Parasitology* **67**: 207-224.

- The *C. elegans* Sequencing Consortium (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**: 2012-2018.
- Voisine, C., Varma, H., Walker, N., Bates, E. A., Stockwell, B. R., and Hart, A. C. (2007) Identification of potential therapeutic drugs for huntington's disease using *Caenorhabditis elegans*. *PLoS One* **2**: e504.
- Waller, P. J., Rudby-Martin, L., Ljungstrom, B. L., and Rydzik, A. (2004) The epidemiology of abomasal nematodes of sheep in Sweden, with particular reference to over-winter survival strategies. *Veterinary Parasitology* **122**: 207-220.
- Waller, P. J., Rydzik, A., Ljungström, B. L., and Törnquist, M. (2006) Towards the eradication of *Haemonchus contortus* from sheep flocks in Sweden. *Veterinary Parasitology* **136**: 367-372.
- Walsh, J. B. and Stephan, W. (2008) Multigene Families: Evolution. eLS doi: 10.1002/9780470015902.a0001702.pub2.
- Whittier, W. D., Zajac, A., and Umberger, S. H. (2009) Control of Internal Parasites in Sheep. *Virginia Cooperative Extension* **410-027**.
- Winston, W. M., Sutherlin, M., Wright, A. J., Feinberg, E. H., and Hunter, C. P. (2007) *Caenorhabditis elegans* SID-2 is required for environmental RNA interference. *Proceedings of the National Academy of Sciences* **104**: 10565-10570.
- Woods, D. J. and Williams, T. M. (2007) The challenges of developing novel antiparasitic drugs. *Invertebrate Neuroscience* **7**: 245-250.
- Xu, S., Liu, C., Tzertzinis, G., Ghedin, E., Evans, C. C., Kaplan, R. *et al.* (2011) In vivo transfection of developmentally competent *Brugia malayi* infective larvae. *International Journal for Parasitology* **41**: 355-362.
- Yatsuda, A. P., Bakker, N., Krijgsveld, J., Knox, D. P., Heck, A. J., and De, V. E. (2006) Identification of secreted cysteine proteases from the parasitic nematode *Haemonchus contortus* detected by biotinylated inhibitors. *Infection and Immunity* **74**: 1989-1993.
- Yin, Y., Martin, J., Abubucker, S., Scott, A. L., McCarter, J. P., Wilson, R. K. *et al.* (2008) Intestinal Transcriptomes of Nematodes: Comparison of the Parasites *Ascaris suum* and *Haemonchus contortus* with the Free-living *Caenorhabditis elegans*. *PLoS Neglected Tropical Diseases* **2**: e269.
- Zhang, B., Pan, X., Cobb, G. P., and Anderson, T. A. (2007) microRNAs as oncogenes and tumor suppressors. *Developmental Biology* **302**: 1-12.
- Zhao, J., Wang, P., and Corsi, A. K. (2007) The *C. elegans* Twist target gene, *arg-1*, is regulated by distinct E box promoter elements. *Mechanisms of Development* **124**: 377-389.

## **Appendices**



## Appendix 1: Common buffers and reagents

|  |  |
|--|--|
| 1x Phosphate Buffered Saline (PBS)           | 137 mM NaCl, 8.1 mM Na <sub>2</sub> HPO <sub>4</sub> , 2.7 mM KCl, 1.47 mM KH <sub>2</sub> PO <sub>4</sub> in sterile distilled H <sub>2</sub> O. pH 7.2. Sterilise by autoclaving, store at room temperature.   |
| 1x Phosphate Buffered Saline Tween-20 (PBST) | Prepare 1 x PBS as above add 1 ml/L Tween-20.  |
| 1x Tris-acetate-EDTA (TAE)                   | 10 mM Tris-HCl pH 8, 1.2 ml acetic acid, 1 mM EDTA pH 8 in distilled H <sub>2</sub> O. Sterilise by autoclaving.   |
| 1x TE  | 10 mM Tris-HCl pH 8, 1 mM EDTA pH 8 in distilled H <sub>2</sub> O. Sterilise by autoclaving.   |
| 2% X-gal                                     | 5-bromo-4-chloro-3-indoyl-β-D-galactosidase (Promega) in N'-dimethyl-formamide. Store at -20°C out of light.   |
| 2x SDS-PAGE buffer                           | 0.09 M Tris-HCl (pH 6.8), 20% Glycerol, 2% SDS, 0.02% bromophenol blue. Store at room temperature.   |
| 5% block solution                            | 10 g skimmed milk powder in 200 ml 1x PBST.  |
| Ampicillin plates                            | L. agar + 1/1,000 dilution 1 M Ampicillin.   |
| Basal media                                  | 100 g bacto-tryptone, 5 g yeast extract, 25 g glucose, 4 g HK <sub>2</sub> PO <sub>4</sub> , 4 g KH <sub>2</sub> PO <sub>4</sub> , 50 ml 250 µg/ml Amphotericin B up to 5 L ddH <sub>2</sub> O. Add 6.5 mg/100 ml Penicillin G, 6.5 mg/100 ml Streptomycin, 25 µl/100 ml Gentamicin, 200 µl 1 M HEPES and 5.2 g Glucose. |
| Blotting buffer                              | 80 ml 10x Tris-glycine solution (BioRad), 160 ml 100% Methanol, 0.8 ml 10% SDS made up to 800 ml H <sub>2</sub> O.   |
| DNA ladder 100 bp and 1 kb                   | 100 µl DNA loading buffer, 50 µl ladder and 350 µl 1x TE.  |
| DNA loading buffer                           | 4 ml 0.5 M EDTA, 25 g FICOL (type 400) and 0.2 g bromophenol blue, made in to 100 ml with H <sub>2</sub> O.  |
| Freezing solution                            | 2.34 g NaCl, 2.72 g KH <sub>2</sub> PO <sub>4</sub> , 120 ml 30% glucose, 2.24 ml 1 M NaOH up to 400 ml ddH <sub>2</sub> O. Sterilise by autoclaving add 1.2 ml 0.1 M MgSO <sub>4</sub> before use.  |
| Glucose Tyrodes balanced salt solution       | 9.6 g Glucose Tyrodes, 1 g NaHCO <sub>3</sub> , 1.2 g Penicillin G, 2 g Steptomycin Sulphate, 0.01 g 5-fluorocytosine in 1 L ddH <sub>2</sub> O.   |
| IPTG   | Isopropyl-β-D-thiogalactoside (Promega) in ddH <sub>2</sub> O. 1 M stock concentration, filter sterilise. Store at -20°C.  |
| L. agar                                      | L. broth + 15 g/L bacto-agar (Oxoid). Sterilise by autoclaving, store at room temperature.   |
| L. broth                                     | 10 g NaCl, 10 g bacto tryptone (Oxoid) and 5 g yeast extract (Oxoid) up to 1 L ddH <sub>2</sub> O. Sterilise by autoclaving, store at room temperature.  |

|                           |   |
|---------------------------|---|
| M9 buffer                 | 3 g $\text{KH}_2\text{PO}_4$ , 6 g $\text{Na}_2\text{HPO}_4$ , 5 g NaCl in 1 L of ddH <sub>2</sub> O. Sterilise by autoclaving add 1 ml $\text{MgSO}_4$ before use.   |
| M9/0.001% Triton solution | 5 $\mu\text{l}$ of 10% Triton X-100 in to 50 ml M9 buffer.  |
| NGM agar plates           | 3 g NaCl, 17 g agar, 2.5 g Peptone in 975 ml of ddH <sub>2</sub> O. Sterilise by autoclaving then add 1 ml of 1 M $\text{CaCl}_2$ , 1 ml of 5 mg/ml cholesterol in ethanol, 1 ml of 1 M $\text{MgSO}_4$ and 25 ml of 1 M $\text{KH}_2\text{PO}_4$ . Store at 4°C.   |
| NGM agar plates + IPTG    | Prepare NGM agar plates as above add 1 ml 1 M IPTG and 1 ml 100 mg/ml Ampicillin. Store at 4°C  |
| Running buffer            | 70 ml 10x Tris-glycine-SDS solution (BioRad) made up to 700 ml H <sub>2</sub> O.  |
| SOC medium                | L. broth + 20 $\mu\text{l}/\text{ml}$ glucose.  |
| SWLB                      | 20 mM Tris pH 7, 50 mM EDTA, 200 mM NaCl and 0.5% SDS. Store at -20°C.  |
| X-gal stain               | 40 $\mu\text{l}$ 1M $\text{NaH}_2\text{PO}_4$ , 210 $\mu\text{l}$ $\text{Na}_2\text{HPO}_4$ , 10 $\mu\text{l}$ 1 M $\text{MgCl}_2$ , 10 $\mu\text{l}$ 0.5 M $\text{K}_4\text{FeCN}_6$ , 10 $\mu\text{l}$ 0.5 M $\text{K}_3\text{FeCN}_6$ , 1 $\mu\text{l}$ 1x SDS and 18 $\mu\text{l}$ 2% X-gal, made up to 1 ml with H <sub>2</sub> O. Made fresh as required. |

## Appendix 2: Primer sequences

### Chapter 3

Table 3.1 5' RACE primer sequences for BAC 18f22 genes

| Name     | 5' → 3' Sequence                  |
|----------|-----------------------------------|
| SJ03_1   | TCTGATAGGCAATCCTCCATCGCACCCGTA    |
| SJ03_2   | AATAAGATCTTGATTCTGATATTTGAACCT    |
| SJ04_1   | TTCGATAGGCCATCCTCCTTCACACCCGTC    |
| SJ04_2   | GAACGTTGTGCAGTTTTTCCAGACGTC       |
| SJ05_1   | TTTGATAGACCAGCCTCCGTCACACCCGAA    |
| SJ05_2   | GAACGAGGTGCAGTTGCTCCAGATTTT       |
| SJ06_1   | TTTGTAGGCCAGCCCCATTACACCCGCT      |
| SJ06_2   | GAGAACAGGGTTTGAACCTGTTTCTCGAA     |
| SJ07_out | AGTCCAGCCTCCTAGACAGTACCCAAGGCCGCA |
| SJ07_in  | AATAGGGTTTATATTGCGGCTTGCGAACTT    |

Table 3.2 *sj02* expression primers giving a fragment of 378 bp on cDNA

| Name          | 5' → 3' Sequence          |
|---------------|---------------------------|
| SJ02_forward1 | CGTCCCTCTGAAGGCTCGAACC    |
| SJ02_exon7    | GGAGAGCAGCATGACAATATGTCTG |

Table 3.3 *sj02* primers for start methionine, 465 bp on cDNA. *sj02* primers in Table 3.2 used as a control.

| Name       | 5' → 3' Sequence            |
|------------|-----------------------------|
| SJ02start  | GTGACCATGAAGTTCTTGGTGTTCACG |
| SJ02_exon7 | GGAGAGCAGCATGACAATATGTCTG   |

Table 3.4 RT-PCR primer sequences for BAC 18f22 genes

| Name          | 5' → 3' Sequence                  |
|---------------|-----------------------------------|
| SJ02_forward1 | CGTCCCTCTGAAGGCTCGAACC            |
| SJ02_exon7    | GGAGAGCAGCATGACAATATGTCTG         |
| SJ03_forward1 | CAAGGATGATTCCGAACCTGTC            |
| SJ03_1        | TCTGATAGGCAATCCTCCATCGCACCCGTA    |
| SJ04_exon4    | GCTACGATCCTCGAGACGTCTGG           |
| SJ04_exon8    | GCACGGTGGGGTTGGCGCTGTTCC          |
| SJ05_forward1 | CATAAAGTTCAGAAATCGGAATC           |
| SJ05_1        | TTTGATAGACCAGCCTCCGTCACACCCGAA    |
| SJ06_forward1 | CGCCTGGTTTTGAGCTCAAATTG           |
| SJ06_1        | TTTGTAGGCCAGCCCCATTACACCCGCT      |
| SJ07_forward1 | GCCGCAATATAAACCTATTGTA            |
| SJ07_out      | AGTCCAGCCTCCTAGACAGTACCCAAGGCCGCA |

## Chapter 4

Table 4.1 RT-PCR primer sequences for *H. contortus* *cpr-6* and *Hc-sod-1* expression

| Name                | 5' → 3' Sequence                     |
|---------------------|--------------------------------------|
| <i>cpr-6_2</i>      | CACACTCAGGCGCACATTATTCACACCCATCAGACC |
| <i>Hccpr6_exon2</i> | CGTACTGGCCGATCGCCGATTGAACTCG         |
| <i>Hcsod-1F1</i>    | CAAAGGCGAAATCAAGGGTTTG               |
| <i>Hcsod-1R1</i>    | AATAACTCCGCAAGCGACAC                 |

Table 4.2 Primer sequences for amplification of 1.2 kb of *H. contortus* *cpr-6* promoter

| Name                | 5' → 3' Sequence                              |
|---------------------|---|
| <i>Hccpr6promF3</i> | GCGCATGCGTGTCTGCGCACAAAGATCATACGGCACTATTTACCC |
| <i>Hccpr6promR</i>  | GCTCTAGACGTGTCCAAGACTGAAGACCACTTTGTTGTGCGGC   |

Table 4.3 Primer sequences for amplification of *C. elegans* *cpr-6*

| Name                | 5' → 3' Sequence                 |
|---------------------|----------------------------------|
| <i>cpr-6GFPFOR2</i> | TATCTGCAGAGTACCATTAACATGCGACAAAC |
| <i>cpr-6GFPREV</i>  | TGACCCGGGAGTAGTTGTCATCGTAGACGTG  |

## Chapter 5

Table 5.1 primer sequences for amplification of a 1 kb *H. contortus* *sj04* promoter region

| Name                | 5' → 3' Sequence                  |
|---------------------|-----------------------------------|
| <i>SJ04forward2</i> | TGCTGCAGCCCACGATCAAACACTGATGTGGG  |
| <i>SJ04reverse1</i> | CGGGATCCATATCTTCAATCTGCCGTTCAATCC |

Table 5.2 *Ce-cpl-1* gene specific primers. Reverse vector primer used with forward gene specific primer as a control

| Name                  | 5' → 3' Sequence                  |
|-----------------------|-----------------------------------|
| <i>T03E6F1</i>        | ACAGCATGCTCCCGAAAAAACTTCAATATTCAG |
| <i>T03E6R1</i>        | CGGTCTAGACTGGAATTTTATAACATTTAAAAT |
| <i>Vector96.04rev</i> | TCTGAGCTCGGTACCCCTCCAAGGG         |

## Appendix 3: Additional figures

Table 1 Additional information for the HmCP genes on scaffold161

| Gene                     | Nembase cluster number   | Percentage DNA identity | Sequence conservation (bp) | Number of ESTs | Published sequences with high identity |
|--------------------------|--------------------------|-------------------------|----------------------------|----------------|--|
| <i>hc-scaffold161-1</i>  | HCC06366_1               | 99                      | 702/1017                   | 15             | EMBL:BF060207                          |
| <i>hc-scaffold161-2</i>  | -                        | -                       | -                          | -              | -                                      |
| <i>hc-scaffold161-3</i>  | HCC03301_1<br>HCC01418_2 | 98<br>98                | 752/932<br>639/932         | 3<br>23        | EMBL:BU665613                          |
| <i>hc-scaffold161-4</i>  | HCC00202_2               | 98                      | 1006/1011                  | 38             | EMBL:Z69345 ( <i>HmCP4</i> )           |
| <i>hc-scaffold161-5</i>  | HCC00328_1               | 98                      | 993/1023                   | 28             | EMBL:Z69344 ( <i>HmCP3</i> )           |
| <i>hc-scaffold161-6</i>  | HCC00473_2               | 99                      | 908/951                    | 53             | EMBL:CA959067                          |
| <i>hc-scaffold161-7</i>  | HCC01418_2               | 99                      | 667/1017                   | 23             | EMBL:CA956449                          |
| <i>hc-scaffold161-8</i>  | HCC00021_1               | 98                      | 956/984                    | 3              | EMBL:Z69346 ( <i>HmCP5</i> )           |
| <i>hc-scaffold161-9</i>  | -                        | -                       | -                          | -              | -                                      |
| <i>hc-scaffold161-10</i> | HCC02966_2               | 96                      | 890/990                    | 136            | EMBL:CA958802                          |
| <i>hc-scaffold161-11</i> | HCC01573_1               | 97                      | 592/855                    | 11             | EMBL:CA869542                          |
| <i>hc-scaffold161-12</i> | -                        | -                       | -                          | -              | -                                      |

Tables 1 and 2 indicate the information obtained from Nembase BLAST searches using the maximum sequence information available for each gene. In addition to EST data, the Washington University BLAST server (<http://www.ebi.ac.uk/Tools/sss/wublast/parasites.html>) was used to identify previously published partial sequences for the genes.

Table 2 Additional information for the *gcp-7*-like genes on scaffold306

| Gene                     | Nembase cluster number | Percentage DNA identity | Sequence conservation (bp) | Number of ESTs | Published sequences with high identity |
|--------------------------|------------------------|-------------------------|----------------------------|----------------|--|
| <i>hc-scaffold306-1</i>  | HCC00358_1             | 96                      | 404/750                    | 5              | EMBL:CB020616                          |
| <i>hc-scaffold306-2</i>  | -                      | -                       | -                          | -              | EMBL:CA958613                          |
| <i>hc-scaffold306-3</i>  | HCC00018_2             | 99                      | 1029/1035                  | 96             | EMBL:Z69342 ( <i>HmCP1</i> )           |
| <i>hc-scaffold306-4</i>  | HCC00019_1             | 97                      | 1013/1035                  | 10             | EMBL:69343 ( <i>HmCP2</i> )            |
| <i>hc-scaffold306-5</i>  | HCC00230_1             | 97                      | 1018/1041                  | 14             | EMBL:BE496731                          |
| <i>hc-scaffold306-6</i>  | HCC01556_1             | 99                      | 1039/1044                  | 9              | EMBL:BG734185                          |
| <i>hc-scaffold306-7</i>  | -                      | -                       | -                          | -              | -                                      |
| <i>hc-scaffold306-8</i>  | HCC00316_1             | 98                      | 532/1044                   | 17             | EMBL:CB099492                          |
| <i>hc-scaffold306-9</i>  | HCC00166_1             | 96                      | 1003/1038                  | 6              | EMBL:AF046229 ( <i>gcp-7</i> )         |
| <i>hc-scaffold306-10</i> | HCC03318_1             | 97                      | 572/1050                   | 3              | EMBL:CA869627                          |
| <i>hc-scaffold306-11</i> | -                      | -                       | -                          | -              | -                                      |
| <i>hc-scaffold306-12</i> | HCC04186_1             | 98                      | 646/1044                   | 3              | EMBL:CA958735                          |
| <i>hc-scaffold306-13</i> | HCC00310_1             | 98                      | 293/690                    | 9              | EMBL:CB015574                          |
| <i>hc-scaffold306-14</i> | HCC00310_2             | 98                      | 742/1035                   | 9              | EMBL:CB015446                          |
| <i>hc-scaffold306-15</i> | -                      | -                       | -                          | -              | -                                      |
| <i>hc-scaffold306-16</i> | HCC01012_1             | 97                      | 1018/1041                  | 6              | EMBL:CA956400                          |
| <i>hc-scaffold306-17</i> | -                      | -                       | -                          | -              | -                                      |
| <i>hc-scaffold306-18</i> | -                      | -                       | -                          | -              | EMBL:CA958026                          |
| <i>hc-scaffold306-19</i> | -                      | -                       | -                          | -              | -                                      |
| <i>hc-scaffold306-20</i> | HCC01173_2             | 95                      | 655/1047                   | 10             | EMBL:CA868860                          |
| <i>hc-scaffold306-21</i> | HCC01124_1             | 95                      | 565/852                    | 3              | EMBL:CB192118                          |