Investigating The Grey Field Slug

WOOD, DOMINIC,MATTHEW

Investigating The Grey Field Slug

Dominic Matthew Wood
2013

Preface

Title: Investigating The Grey Field Slug: Investigation, Analysis and Application of transcriptomic data of crop pest *Deroceras reticulatum*

Author: Dominic Matthew Wood
Email: dominic.matthew.wood@gmail.com
Department: School of Biological & Biomedical Sciences
Institution: University of Durham

Date: August 2013

Declaration

No material here has previously been submitted for any other degree. Except where acknowledged, all material is the work of the author.

Statement of Copyright

Date……
Signed……

Abstract

High-throughput sequencing was used to analyse cDNA generated from tissues of the grey field slug, *Deroceras reticulatum*, a significant invertebrate pest of agricultural and horticultural crops. Almost no sequence data is available for this organism. In this project, we performed *de novo* transcriptome sequencing to produce sequence dataset for the *Deroceras reticulatum*.

A total of 132,597 and 161,419 sequencing reads between 50-600bp from the digestive gland and neural tissue were obtained through Roche 454 pyrosequencing. These reads were assembled into contiguous sequences and annotated using sequence homology search tools. Multiple sequence assemblies and annotation data was amalgamated into a biological database using BioSQL. Analysis of the dataset with predictions of probable protein function were made based on annotation data. InterPro (IPR) terms generated with InterProScan software were mapped to read counts and used to identify more frequently sequenced gene families.

Digestive hydrolases were major transcripts in the digestive gland, with cysteine proteinases and cellulases being the most abundant functional classes. A Cathepsin L homologue is likely to be responsible for the proteinase activity of the digestive gland which was previously detected by biochemical analysis. Cathepsin L and several other predicted proteins were used to design RNAi experiments to assess potential for crop pest defence strategy. Further work on protein expression of a native tumour necrosis factor (TNF) ligand homologue was also conducted as an exemplar study.

Contents

Preface 2

Abstract 3

Contents 4

List of Tables 9

List of Figures 11

Acknowledgements 14

Chapter 1 | Introduction 15

1.1 Introduction to Mollusca and the Target Species, *Deroceras reticulatum*, the Grey Field Slug 15

1.1.1 *Deroceras reticulatum* 16
1.1.2 *Deroceras reticulatum* digestive gland 18
1.1.3 *Deroceras reticulatum* nervous system 21
1.2 Mollusca as Crop Pests 22

1.2.1 Current Molluscicide Usage 24
1.2.2 Metaldehyde and Methiocarb 25
1.2.3 Other molluscicides 26
1.2.4 Biological impact of molluscicides 28
1.3 Mollusca at the Molecular Level; Current Molluscan Genetic data 29

1.3.1 Pyrosequencing 29
1.3.2 Sequence Data Assembly 31
1.3.3 Transcriptomics and Identifying Biochemistry 33
1.3.4 Biological Databases and BioSQL 34
1.4 RNA Interference and Crop Protection against Molluscs 38

1.4.1 RNAi against Mollusca 41
1.5 Aims and Objectives of the Project 43

Chapter 2 | Materials & Methods 45

Contents

2.1 General molecular biology methods 45
2.1.1 Recipes 45
2.1.2 Phenol-chloroform ethanol precipitation 47
2.1.3 Oligonucleotides 48
2.1.3 Degenerate Primer Design 48
2.1.2 DNA amplification with Polymerase Chain Reaction (PCR) 48
2.1.3 High Fidelity PCR 49
2.1.4 Touch Down PCR Protocol for degenerate primers 49
2.1.5 Colony PCR 49
2.1.6 Nucleic Acid Quantification 50
2.1.6 DNA ligations 50
2.1.7 Isolation of plasmid DNA 50
2.1.8 Restriction endonuclease digestion of DNA 51
2.1.9 Agarose Gel electrophoresis 51
2.1.10 Gel Extraction of DNA 51
2.1.11 Agarose Gel electrophoresis for size separation of cDNA 52
2.1.12 SDS-PAGE electrophoresis 52
2.1.13 Staining with Coomassie Brilliant Blue 52
2.1.14 Western Blotting 52
2.1.15 Chemiluminescent detection of membranes 53
2.1.16 Glycerol stocks of *E. coli* strains 53
2.1.17 *E. coli t*ransformation 54
2.2 RNA and cDNA 54
2.2.1 Preparation of digestive gland total RNA for SMART cDNA 54
2.2.2 Preparation of neuronal tissue for SMART cDNA 55
2.2.3 mRNA enrichment of total RNA 55
2.2.4 SMART cDNA synthesis for high throughput pyrosequencing 55
2.2.5 Rapid Amplification of cDNA Ends (RACE) 56
2.2.6 cDNA synthesis with random hexamers 56
2.3 Molecular cloning 56
2.3.1 Cloning cathepsin L dsRNA construct 58
2.3.2 Cloning apoptosis dsRNA construct 59
2.3.3 Cloning GAPDH dsRNA constructs 60

2.3.4 Cloning dTNF protein expression construct 60
2.3.5 Site directed mutagenesis of dTNF 61
2.4 dsRNA Production 62
2.4.1 *in-vitro* dsRNA Production 62
2.4.2 *in-vivo* dsRNA production 62
2.5 Quantitative PCR methods 63
2.5.1 Total RNA extraction for qPCR 63
2.5.2 Quantitative PCR (qPCR) 64
2.5.3 qPCR Analysis 64
2.5.4 HKG analysis 64
2.6 Protein Expression Methods 65
2.6.1 Calculating Protein Molecular Mass 65
2.6.2 Protein Extraction Method 65
2.6.3 Protein Purification with HisTrap column 66
2.6.4 Dialysis & lypholization 66
2.6.5 Buffer Exchange Column Purification 67
2.6.6 BCA Assay 67
2.7 *Deroceras reticulatum* Methods 67
2.7.1 *Deroceras reticulatum* sourcing 67
2.7.2 Maintaining cultures 68
2.7.3 Injection Assays 69
2.8 Bioinformatic Methods 69
2.8.1 Sequencing 69
2.8.2 Univec Cleanup 70
2.8.3 Sequence Data Assembly 70
2.8.4 Dataset Upload 70
2.8.5 BLAST Homology 71
2.8.6 Peptide Prediction & InterProScan 71
2.8.7 Phylogenetic Analysis 71
2.9 Biology Database 72
2.9.1 Data upload 72
2.9.2 Taxonomy Mapping 73

2.9.3 Assembly & Homology Metrics 73
Chapter 3 | Initial Investigation of Transcripts from *D. reticulatum* 77
3.1 Extraction of RNA and cDNA synthesis 79
3.1.1 Extraction of RNA from *D. reticulatum* 79
3.1.2 Random cloning into pJET2.1 vector 82
3.2 Characterisation of full-length transcripts; 5' & 3' RACE with Ferritin Homologue 84
3.3 Ion Channels as a Target for Molluscicides; Degenerate Primer PCR 87
3.3.1 Partial sodium channel sequence 88
3.4 High Throughput Sequencing 93
3.4.1 454 Sequencing 93
3.4.2 Sequence Cleaning with Univec 93
3.4.3 Assembling Digestive Gland Data 94
3.4.4 Assembling Neuronal Tissue Data 95
3.4.5 Comparison of Assemblers 98
3.4.6 Contig Naming 99
Chapter 4 | Analysis: Transcript Sequences 100
4.1 BLASTx homology analysis of Digestive Gland transcriptome data 100
4.1.1 Species and Phyla of homology matches 102
4.2 ESTScan peptide prediction and InterProScan 106
4.2.1 Analysis of IPR terms for Digestive Gland tissue 108
4.3 Largest protein groups based on IPR terms 110
4.3.1 Cysteine Peptidase 110
4.2.2 Ferritin 115
4.3.3 Glycosidase Hydrolase Family 9 119
4.4 Other notable groups 123
4.4.1 C-Type Lectins 123
4.4.2 P60-like 127
4.4.3 Other Glycoside Hydrolases 127
4.5 BLAST homology analysis of Neural tissue data 134

4.5.1 INTERPRO Analysis of Neuronal Tissue 136
4.6 Notable Transcripts in Neuronal Tissue 139
4.6.1 Tumour Necrosis Factor Like Proteins 140
4.6.2 Vasopressin Like Peptide 143
4.6.3 Spider Toxin-Like Protein 145
Chapter 5 | Application: Targeting *D. reticulatum* 150
5.1 RNAi against Cathepsin L 150
5.1.1 dsRNA Production and Injection Assay for RNAi Effects 151
5.1.2 Quantitative qPCR on Cathepsin L RNAi injections 156
5.2 Assay of Normalisation Standards for qPCR (HKGs) for *D. reticulatum* 159
5.3 Down Regulation of Expression of Apoptosis Inhibitor by RNAi 163
5.3.1 RNAi Construct & Assay 165
5.4 RNAi against GAPDH 168
5.4.1 *GAPDH* dsRNA Injections at 20µg 170
5.4.2 GAPDH dsRNA Injections at 50µg with *in-vivo* produced dsRNA 170
5.5 TNF Ligand-like gene as a molluscicidal agent 179
5.5.1 Recombinant Protein Expression 179
5.5.2 Injection Assay 183
Chapter 6 | Discussion 185
6.1 Assembly Comparisons 185
6.2 Use of contigs, reads and coverage 186
6.3 Homology Analysis Caveats 187
6.4 Programming Tools 188
6.5 General Conclusion 189
Bibliography 195

List of Tables

Table 1 : Example of output from an SQL query on the BioSQL database 37
Table 2: Cathepsin L RNAi construct primers 58
Table 3: Apoptosis Inhibitor RNAi Construct and qPCR primers 59
Table 4: GAPDH RNAi construct primers 60
Table 5: dTNF cloning and mutagenesis primers 61
Table 6: Primers for housekeeping gene qPCR primers 65
Table 7: pJET digestive gland sequenced clones and BLAST database matches 83
Table 8: Redundant Primers designed based on A. Californica Na voltage gated channel 88
Table 9: Table of Gene specific primers for D. reticulatum Sodium Channel Fragments 91
Table 10: Table showing statistics for Univec screen 94
Table 11: Statistics of 3 Assemblies of Digestive Gland 454 Read Sequences 95
Table 12: Statistics of 3 Assemblies of Neuronal Tissue 454 Read Sequences 97
Table 13: Summary of BLAST homology matches for digestive gland assemblies 102
Table 14: Summary of top IPR terms 107
Table 15: Summary of IPR terms only significant in a single assembly 108
Table 16: Top active sites based on IPR terms 110
Table 17: Contigs with homology to cysteine peptidase like genes 111
Table 18: Table of all glycosidase family IPR terms represented in the digestive tissue data 129
Table 19: Summary of BLAST homology matches for neural tissue assemblies ... 134
Table 20: Summary of top IPR terms for neural tissue assemblies 139
Table 21 : T-test values of Cathepsin L RNAi experiment 157
Table 22: Table of 5 HKG analysed for stability using qPCR Ct value data 160
Table 23: HKG analysis of 3 HKG with digestive gland and whole tissue qPCR data 161
Table 24: HKG analysis of 3 HKG with whole tissue qPCR data 162
Table 25: Table showing Ct values of 3 HKG in digestive gland and whole tissue 163

List of Tables

Table 26: Ct values for two qPCR studies on RNAi data 176
Table 27: Comparison of Biological Replicates for GAPDH RNAi 177
Table 28: Comparison of HKG Replicates for GAPDH RNAi 177
Table 29 : Independent T-Test of Variables for GAPDH RNAi 178

List of Figures

Figure 1 : Alimentary tract and accessory organs of Deroceras reticulatum 20
Figure 2 : Change in treatment area of molluscicides and total pesticides 23
Figure 3 : Simple example of a relational database model 35
Figure 4 : General stages of RNA interference 40
Figure 5 : Deroceras reticulatum 68
Figure 7 : Example Total RNA and cDNA 78
Figure 8 : Gel Electrophoresis of Total RNA with the two main rRNA subunits 80
Figure 9 : Gel electrophoresis of D. reticulatum cDNA samples 81
Figure 10 : Gel electrophoresis of RACE PCR products for Ferritin Gene 85
Figure 11 : D. reticulatum sequence homologous to Ferritin identified through 5' & 3' RACE 86
Figure 12 : Structure of A. californica Ion channel with primer locations flagged. .. 87
Figure 12 : Structure of A. californica Ion channel with primer locations flagged. .. 87
Figure 13 : Gel electrophoresis of redundant sodium channel PCR products 89
Figure 14 : Alignment of D. reticulatum cDNA sequence with A. californica SCAP1 gene 90
Figure 15 : Alignment of D. reticulatum protein sequence with A. californica SCAP1 gene 91
Figure 16 : Kernel Density plot of 454 reads for Neural and Digestive tissue sources 92
Figure 17 : Kernel density plot of digestive gland contig lengths 96
Figure 18 : Kernel density plot of neural tissue contig lengths 98
Figure 19 : Number of top BLAST hits below BLAST evalue for 3 assemblies 101
Figure 20 : Venn diagram of reads with BLAST matches below 1e-3 103
Figure 21 : Top BLAST matches per phyla as a percentage for the digestive gland 104
Figure 22 : Breakdown of top hits by species for the digestive gland assemblies .. 105
Figure 23 : Cathepsin L homologue of D. reticulatum alignment 112
Figure 24 : Phylogenetic tree of the cathepsin L from Deroceras reticulatum 113

List of Figures

Figure 25 : Alignment of contig with serine protease homologues 116
Figure 26 : D. reticulatum contig homologous to L. stagnalis snail soma ferritin .. 117
Figure 27 : SIREs RNA fold prediction for D. reticulatum Ferritin sequence 118
Figure 28 : Alignment of D. reticulatum contig with Molluscan homologues..... 120
Figure 29 : Phylogenetic tree of D. reticulatum GHF9 with homologues 122
Figure 30 : Sequence alignment of C-Type lectins..... 124
Figure 31: Phylogenetic tree of Deroceras reticulatum and C-Type Lectin homologues 126
Figure 32 : Alignment of D. reticulatum contig with GHF13 homologues..... 132
Figure 33 : Number of top BLAST hits below BLAST evalue for 3 assemblies 135
Figure 34 : Top BLAST matches per phyla as a percentage for the neural tissue... 136
Figure 35 : Breakdown of top hits by species for the neural tissue assemblies..... 137
Figure 36 : Alignment of smaller TNF domain containing peptides 140
Figure 37 : Alignment of C1qDC peptide with homologue from C. gigas 141
Figure 38 : Alignment of D. reticulatum TNF ligand like peptide with homologue 143
Figure 39 : Comparison of D. reticulatum conopressin like contig with homologues 144
Figure 40 : Alignment of two D. reticulatum peptides with spider toxins and knottin containing insect peptides 146
Figure 41 : Phylogenetic Tree based on the alignment of cysteine knot domains... 148
Figure 42 : PCR products Cathepsin 5' and 3' Fragments for RNAi Constructs 152
Figure 43 : Dye injection assay..... 153
Figure 44 : Cathepsin L RNAi fragments 154
Figure 45 : ssRNA for Kanamycin, 5’ & 3’ Cath and Apin RNAi experiments 155
Figure 46 : in-vitro transcribed dsRNA of Kanamycin and Cathepsin L fragments 156
Figure 47 : Cathepsin Levels for individual organisms in RNAi Study 158
Figure 48 : Average RQ values for RNAi against Cathepsin L..... 159
Figure 49 : Comparison of D. reticulatum contig with A. californica apoptosis inhibitor..... 164
Figure 50 : Apoptosis Inhibitor RNAi fragment..... 165

List of Figures

Figure 51 : D. reticulatum Apoptosis Inhibitor PCR and dsRNA electrophoresis gels 166
Figure 52 : GAPDH RNAi fragments..... 168
Figure 53 : GAPDH PCR product for insertion into RNAi transcription vector..... 168
Figure 54 : ssRNA and dsRNA synthesised through in-vitro transcription of GAPDH RNAi Construct 169
Figure 55 : Relative quantitation for GAPDH RNAi study at 20µg..... 171
Figure 56 : GAPDH in vivo dsRNA production..... 172
Figure 57 : Relative quantitation of GAPDH vs. Actin and EF1-A 173
Figure 58: Linear Regression of GAPDH VS HKG..... 175
Figure 59 : Protein sequence of Trx-dTNF with highlighted regions..... 180
Figure 60 : CBB and Western blots showing purification steps for TRX-dTNF 181
Figure 61 : Coomassie brilliant blue and Western blots of buffer exchange purified dTNF..... 183

Acknowledgements

I am very greatful for the opportunity awarded to me by Prof. John A. Gatehouse who has always been strongly supportive. From the start of my PhD he has always treated me as an academic equal and I appreciate his reasoned thinking and open mind. Many thanks to the members of Lab 1 Research group, in particular Dr. Catherine Bruce who was my intial mentor as well as the excellent academic expertise and friendship of Dr. Prashant Pyati.

Thanks to the past and present members of lab 1 I worked with, who created a friendly working environment, and offered endless advice. This included Dr. Elaine Fitches, Dr. Dan Price, Dr. Emma Back, Dr. Gareth Hinchliffe, Min Cao, Sheung Yang and all other members of Lab 1 and members of the school of Biological and Biomedical Sciences who have aided me throughout my time at Durham University.

Additionally I would like to thank Dr Richard Thwaites and his team at the Food and Environment Research Agency who conducted the 454 sequencing that made this project possible.

Chapter 1 | Introduction

1.1 Introduction to Mollusca and the Target Species, *Deroceras reticulatum*, the Grey Field Slug

The Mollusca phylum is one of the largest and most diverse phyla in size and variety, second only to insects. They are the largest marine phylum, but are also found in freshwater and terrestrial habitats, and include familiar pests such as slugs and snails, as well as marine organisms such as clams and abalone. Mollusc species range in size from micromolluscs, including the gastropod family *Omalogyridae* (less than 1mm in length) to the giant and colossal squids (*Architeuthis dux* and *Mesonychoteuthis hamiltoni* respectively), which can grow up to 14m in length. The first mollusc-like creatures are generally thought to have evolved approximately 500 million years ago during the Cambrian explosion, a rapid increase in animal species with the appearance of most of the present day metazoan phyla. The early molluscs were protected by a cuticle of aragonitic spicules or scales rather than the shells that later evolved in modern molluscs (Scheltema and Schander 2006). Several mollusc classes then went through repeated and independent evolutions whereby shells were lost through reduction and internalisation (Osterauer et al. 2010). In the Gastropoda class this lead to the evolution of snail and slug species such as *Deroceras reticulatum*.

The general anatomy of molluscs is that of a body divided into two: a visceral mass containing most of the organs and a combined head and foot. A shell, secreted by the mantle (a specialised area of the body wall) covers the visceral mass in many shellfish and snail species, but is internalised in other molluscs such as slugs, octopuses and squid. There is no supporting endo- or exo- skeleton although in many cases the shell performs this function. The body is not segmented and the body cavity filled with heamocoel. Mucus is secreted from the skin coating the body surface. The feeding organ is a hardened spiked tongue, called a radula, which rasps at food during feeding. The nervous system is ganglionated, and is a model neurological

system with the giant squid axon biology being key to uncovering ionic mechanism of action potentials (Hodgkin and Huxley 1952).

The repeated evolution of molluscs that conform to the general appearance of slugs and snails mean that the terms are only referent to body type, and are not a distinct taxonomic group. However the majority of air-breathing terrestrial slugs and snails belong to the taxonomic group Stylommataophora, which is considered a clade of Gastropoda. Although snails, with an external shell to protect them from moisture loss, might appear to be better adapted for terrestrial environments, the internalisation of external calcified shells in terrestrial slug species allows them to access smaller spaces. The evolution of this trait is thought to provide better adaptation to soil as a habitation for slug species (Barker 2001). Additionally the lesser need for supplies of calcium to synthesise the calcareous shells allows slugs to thrive in a wider range of soil types and qualities than snail counterparts. Of the 95 Stylommataophora classified by Boycott, only 1 of the 37 snail species was able to tolerate low calcium soils, living in acid heath and woodland areas. All 58 slug species showed indifference to soil type and as such had a larger range of habitation (Boycott 1934). However, due to a reduced capacity to tolerate dry conditions, as a result of losing the protection of an outer shell, the distribution of slugs is restricted to areas which are generally high in rainfall (Solem 1974). This has not stopped slug species becoming a worldwide pest to many agricultural and horticultural crops, one of the most notable being *D. reticulatum*.

1.1.1 *Deroceras reticulatum*

D. reticulatum was first described as *Limax reticulatus* by the Danish naturalist Otto Friedrich Müller in 1774. Since then its scientific name has changed, with genus changing from *Limax* to *Agriolimax* to the present day consensus of *Deroceras*, with the additional alteration of *reticulatus* to *reticulatum*. Its common name also has synonymous variations including the grey field slug, grey garden slug and netted slug. This species is one of the most common terrestrial molluscs in Northern Europe, with global distribution across most temperate and sub-tropical regions including North and South America (Tulli et al. 2009) and Australasia

(Ferguson, Barratt, and Jones 1988). They are primarily synanthropic, inhabiting cultivated areas such as arable land but extending to fields, meadows, gardens and parks. They are not commonly found in woodland or forests. Slugs, in particular *D. reticulatum*, are a major pest of many crops in the UK and Northern Europe, with winter wheat being one of the most important economically. In severe cases, when weather conditions are particularly favourable for slugs, they can cause damage to up to a third of seeds and seedlings in autumn. Their significance as a pest species has increased over the last 30 years, primarily attributed to changing agricultural practices (Brooks et al. 2003).

Individual adult *D. reticulatum* grow to 3.5-5 cm in length when fully extended and can range in colour from dark-brown to pale cream. Initial recently hatched juveniles have a glassy transparent appearance, with the internal organ structure visible. As the slug reaches adulthood a mottled, 'netted' pattern develops and becomes more distinct; the body surface darkens, and a particular fingerprint-like pattern develops on the mantle. As with all terrestrial slugs, mucus is exuded from the body surface, the surface being ridged with tubercles giving it a grooved appearance. When the slug is agitated the mucus, which is normally clear and colourless, becomes milky white and viscous. Additionally, the agitated slug can contract its head by drawing it in anteriorly, with mantle forming a flap that covers the head and neck.

Whilst the growth, metabolism and reproduction cycles of *D. reticulatum* and other terrestrial slug pests of crops are much slower than equivalent insect pests, their populations rapidly regenerate despite agricultural counter-measures. *D. reticulatum* is a protandric hermaphrodite, whereby male sexual mature phase is reached early in the lifespan followed by female, with substantial overlap between the two phases (Runham 1978). Sexual reproduction is relatively complex with courtship and copulation including relatively elaborate stages such as trail following, pairing and circling (Nicholas 1984). Despite self-fertilisation being common in many slug species *D. reticulatum* rarely produces self-fertilised eggs, relying on them only as stop-gap measure when populations are low, with resulting offspring having poorer survival and fecundity (Howlett 2005). A single *D. reticulatum* can lay 500 eggs with

batches of eggs being laid throughout the year, although their peaks egg laying periods are based on climate and can shift between spring and autumn depending on year to year conditions (Runham and Hunter 1970).

D. reticulatum live within the soil, primarily within the top 10 cm, and feed mainly at night (Glen and Symondson 2002). *D. reticulatum* are particularly resistant to cold conditions compared to other slug species and feed normally at temperatures as low as 0°C (Mellanby 1961). During very cold periods *D. reticulatum* remain dormant in the soil and can survive for more than 3 months without any food (Middlebrooks, Pierce, and Bell 2011). *D. reticulatum* as with many terrestrial molluscs feeds on and digests a wide range of foodstuffs.

D. reticulatum is an omnivorous and a generalist grazer and its specific interaction with plant species is less direct than cereal pests such as *Sitobion avenae* and *Mayetiola destructor*. As yet there are few examples of resistance genes, in plants, to molluscs, equivalent to the gene-for-gene interactions described between insects and plants. Plant populations have been shown to be affected by molluscan herbivory with snail herbivory affecting secondary metabolites concentrations in willow plants. However the resulting plants were more susceptible to other herbivores and rust fungus, highlighting that molluscs are part of a more complex herbivorous population (Orians et al., 2013). The current research suggests plants rely on general host-plant resistance (HPR) genes that affect a wider variety of herbivores rather than targeted genes for mollusc resistance. Secondary metabolites such as glucosinolates have shown to negatively affect multiple generalist species including snail species (Newton et al., 2009), with comparison on mollusc grazing of seawood showing plant defences preferentially reducing generalist mollusc species over specialists (Long et al., 2007).

1.1.2 *Deroceras reticulatum* digestive gland

The alimentary tract of *D. reticulatum* can be split into 6 regions; buccal mass, oesophagus, crop, stomach, intestine and rectum. Food sources such as plant leaves are broken by radula with the rasping action aided by the jaw. When feeding the animal conducts multiple rasps whilst moving the head side to side and slowly

moving forward. The teeth are significantly worn down by this action, with worn teeth at the front of the radula which have become rounded to stumps falling off; the cast-off teeth are then eaten by the organism. The rate of replacement of rows of teeth on the radula was found to vary among species with mature *D. reticulatum* replacing 5.6 rows of teeth per day (Isarankura and Runham 1968).

Food particles are then transferred through the buccal cavity, where they are combined with saliva, a combination of mucus and digestion-related enzymes. The food is quickly transported to the crop via the oesophagus, which has a ciliated epithelium as well as being lubricated by mucus to allow transport of the food particles. The crop contains a viscous brown liquid containing a much larger complement of enzymes than the saliva. Food remains in the crop whilst extracellular digestion proceeds aided by circulation of food particles by peristaltic contractions of the crop wall. Remaining food matter continues through into the stomach, and faecal matter begins to form as food remains are compacted and moisture is extracted in the intestine. Mucus is added along the length of the intestine and faeces are discharged from the body as a faecal string.

There are two main types of accessory glands which produce digestive enzymes, the salivary glands and digestive gland. The left and right salivary glands are tear shaped organs near the nerve ring at the border of the oesophagus and crop and are connected to the buccal mass via ducts. The digestive gland is a large multilobed structure, the lobes of which wrap around the stomach and intestine of the alimentary tract, see Figure 1. (Walker 1972). The digestive gland is responsible for 82% of proteolytic activity in the gut and crop with salivary glands being less important, contributing 13% and 5% of activity in the gut and crop respectively. In *D. reticulatum* crop digestive juice was demonstrated to contain activities of amylase, cellulase, α-glycosidase (invertase), xylanase, chitinase, gelatinase and lipase (Runham and Hunter 1970). A number of different protease activities were also described, distinguished from commensal digestion through antibiotic treatment assays. The primary protease, identified as a cysteine protease (likely Cathepsin L), was responsible for the majority of proteolysis in the digestive gland (Walker et al., 1998). Carbohydrase, lipase and proteases are expected components of digestive

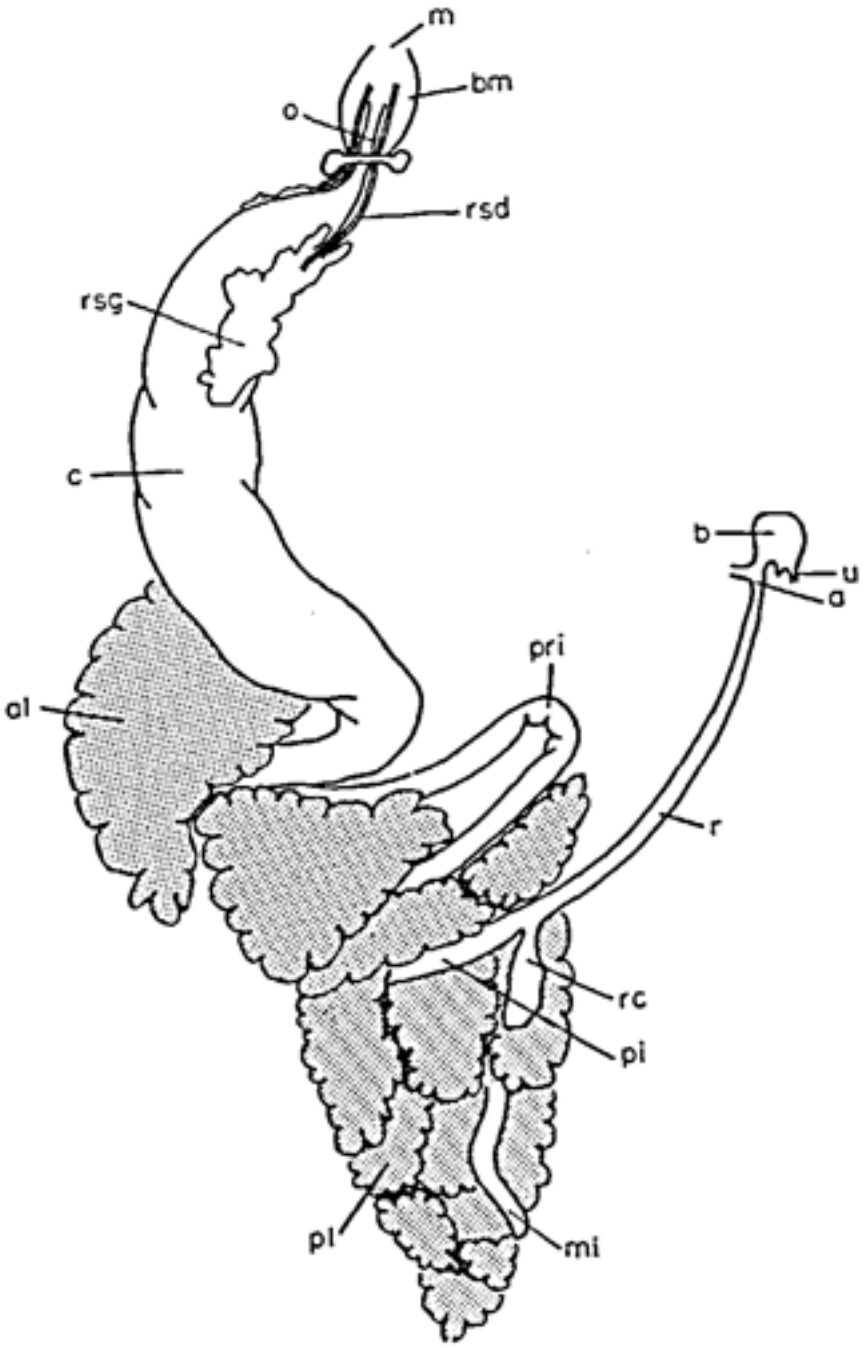


Figure 1 : Alimentary tract and accessory organs of **Deroceras reticulatum**
a, Anus; al, anterior lobes of digestive gland; b, bladder; bm, buccal mass; c, crop; m, mouth; mi, mid-intestine; o, oesophagus; pi, post-intestine; pl, posterior lobes of digestive gland; pri, pro-intestinal loop; r, rectum; rc, rectal caecum; rsd, right salivary gland; u, ureter.

systems which hydrolise plant matter into monomers. The presence of chitinase could be explained due to feeding on insects as well as the recycling of worn teeth which contain chitin (Isarankura and Runham 1968; Sollas 1907).

1.1.3 *Deroceras reticulatum* nervous system

The 'brain' of *D. reticulatum* is a ring shaped complex association of ganglia which encircles the oesophagus and salivary gland ducts. Around the outside of the ring of ganglia is a thick sheath containing collagen, muscle fibres and various connective tissue cells, some of which appear to be secretory. Although secretion by nerve cells is generally short-range, in the form of neurotransmitters which are released at the synapse, travelling short distances in order to stimulate nearby cells. Neurosecretions can also be released into the blood or connective tissue of a group of cells at a distance. Nerves run out from the nerve ring in close association with arteries, often bundled together in the same sheath. This includes the optic nerve which runs along the optic tentacles and connects to the retina. In all sensory tentacles the tip contains a large digitate ganglion with a retractor muscle attached around its edges. These ganglia are connected to the nerve cord with an olfactory nerve and are responsible for the organism's sense of smell (Runham and Hunter 1970).

Beyond the basic anatomy, the *D. reticulatum* nervous system has not been studied in any detail. However, the nervous systems of many other molluscs have been studied extensively. *Aplysia californica*, a marine Gastropoda, is a model organism for neurobiology and has received considerable attention, including a large database of expressed sequence tags (ESTs) and a draft genome is in progress. Many genes of significance to the nervous system, including those encoding channels, receptors, and hormones, have been sequenced and predicted protein sequences have been produced (Sattelle and Buckingham 2006). These in many cases are a starting point for work examining the genetic aspects of the *D. reticulatum* nervous system.

1.2 Mollusca as Crop Pests

Molluscan crop pests reduce seedling growth, and sever reproductive and vegetative tillers affecting cereal crop yields. Slug grazing reduces overall leaf biomass, and fouls flowers with mucus, reducing pollination through a decrease in attractivity to bees (*Apis mellifera*) (Gavin et al. 2007); severe slug grazing can also affect subsequent crops due to the buildup in population which persists from one growing season to another. Slugs devastate newly planted cereal fields by hollowing out seeds and grazing on seedlings, most seriously in wheat (*Triticum aestivum*). Wheat is a key cereal crop for a large proportion of the globe and is the largest single agricultural product of Europe, with 225 million metric tons produced in 2011 (FAOSTAT United Nations 2011). Increasing demand due to population increase, has put ever more pressure on reducing crop wastage due to pest damage, possible use of crops to generate biofuels and projected effects of climate variability reinforce the need to maximise yields of harvested products.

Besides causing yield losses, slugs can have major impacts on acceptability of crop products to consumers. For example, due to the increasing demand for fresher foodstuffs, pea vining is now frequently done at night, when slugs are most active and highest up within the crop, resulting in slugs being co-harvested with the peas. The harvest is sent to freezing plants, where the separation of frozen slugs and pea seeds causes great difficulty (Runham and Hunter 1970). Additionally, other crops which are grown in much smaller quantities but have a greater commercial value, such as strawberries, asparagus and salad vegetables are particularly susceptible to slug damage. In these cases minor grazing and cosmetic damage such as slug trails and faeces renders crops unmarketable or of very little value. Consumers and supermarkets are increasingly unwilling to tolerate cosmetic damage whilst also demanding fresher food, leading to greater demand for crop protection mechanisms against molluscs. With new markets such as 'organic' foodstuffs gaining market share, producers are also looking for alternatives which may be more acceptable to consumers than chemical- based pesticides (Glen 2002).

In many cases crop rotations and level of soil disruption through tillage can have a pronounced effect on slug populations. However variability in these

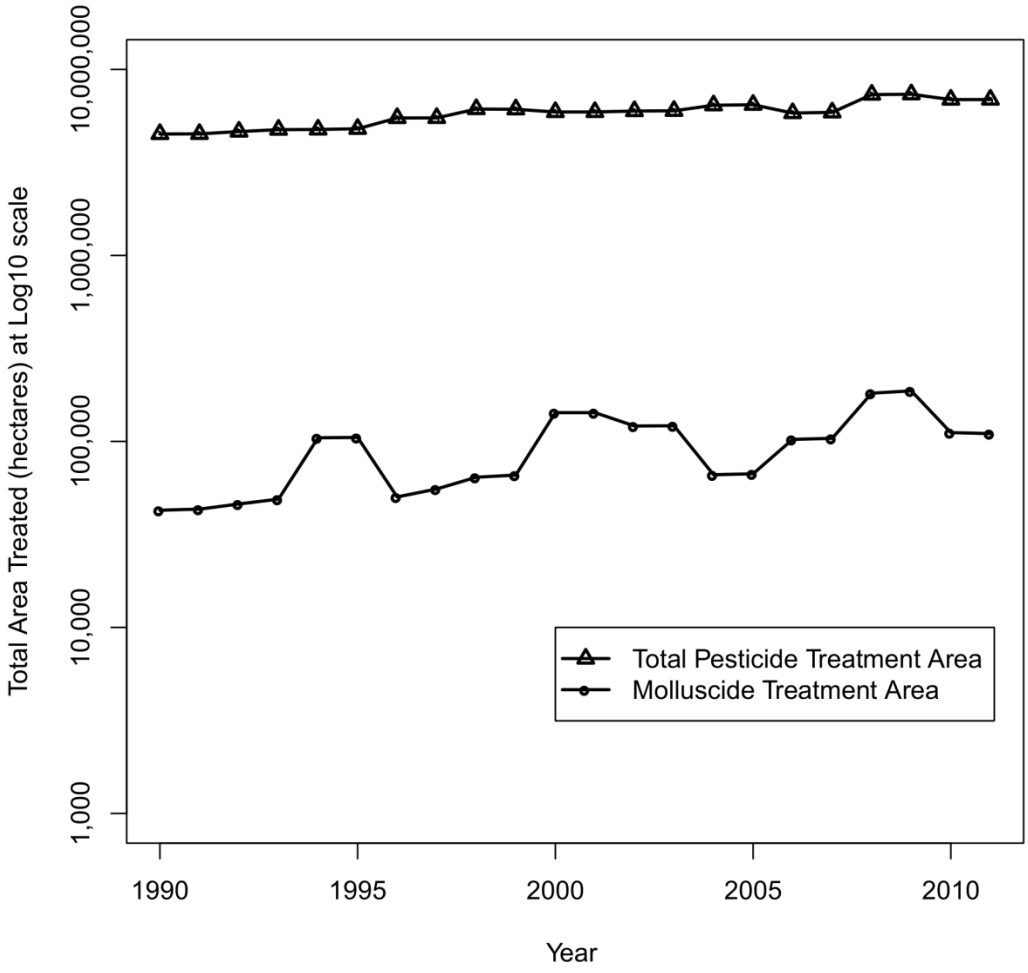


Figure 2 : Change in treatment area of molluscicides and total pesticides

Graph shows the overall trends of molluscicide usage measured by treatment area, based on FERA pesticide survey data. Total pesticide usage, which includes molluscicides, has been included for reference. Both values show on overall increasing trend over the 2 decades that data is available.*

**Surveys for some pesticide groups are not conducted each year, values published by FERA are sometimes interpolated from surrounding years.*

populations through crop type and weather from year to year mean it is difficult to effectively protect crops through soil management alone. Many modern crop

practices have been shown to promote slug populations. No-till farming is increasingly popular method of farming which implements a method of continuously growing crops all year round without deep ploughing or major disruption of the soil. This style of farming creates long-term habitats for slugs as well as reducing damage to slug populations caused by soil tilling (Glen and Symondson 2002).

Seed drilling methods, the process of planting seeds into the soil, are another area of change in agricultural practice which has given rise to greater potential damage by slugs. Traditionally, seed drilling was done with a much larger number of seeds than required, and unwanted plants were removed after seeds had sprouted. Modern agricultural technology now allows for precision monogerm drilling of single seeds, which reduces overall waste. However, no adjacent plants are available if the sprouting seed is damaged, increasing the risk of having to replant entire crops when slug populations are significant (Runham and Hunter 1970). Currently the primary strategy for protecting most crop types is the use of molluscicides, but these can show variable results depending on how and when they are used.

1.2.1 Current Molluscicide Usage

The Pesticide Usage Survey conducted on UK arable crops by the Department of Food, Environment and Rural Affairs (FERA), showed during the first decade of the 21st century, on average 1/4 of UK arable land area was treated with molluscicidal agents. 95% of these molluscicides are used on 3 main crops: wheat (52%), oilseed rape (33%) and ware potatoes (10%) (Garthwaite et al. 2010). Whilst the latest FERA survey reported decreased usage in 2010 and 2011, a broader look at the usage data indicates a quite high variability year on year for molluscicides, see Figure 2. This may be more likely due to climate variation having impact both on mollusc populations and the efficacy of the molluscicides under different climatic conditions. In autumn 2012 the Metaldehyde Stewardship Group, a consortium of metaldehyde suppliers, reported depleted stocks of slug pellets as well as ‘highly concerning’ metaldehyde levels in drinking water. Due to wet weather and lack of severe frosts or very cold weather, unprecedented slug pressure forced multiple applications of slug pellets to avoid total crop devastation (“Metaldehyde

Stewardship Group 22/10/12 - Briefing Notes: Autumn 2012" 2012). The general trend for both total pesticides and molluscicides over the last 20 years is an increase in total amount used.

1.2.2 Metaldehyde and Methiocarb

Two main agents, metaldehyde and carbamates, primarily methiocarb, represent 99% of molluscicides used by the UK agricultural industry. Carbamate pesticides are biologically active because they are complementary in structure to the active site of acetylcholinesterase. They behave as synthetic neurohormones that cause toxic action by interrupting the normal action of acetylcholinesterase so that acetylcholine accumulates at synaptic junctions (Metcalf 1971). This leads to interference of the metabolism of the gastropod nervous system, resulting in mortality (Godan 1999). Metaldehyde is a cyclic tetramer of acetaldehyde, with which it shares many chemical properties. Its mode of action is less well characterised than carbamates. Initially it was thought that acetaldehyde, created through hydrolysis of metaldehydes, was the primary toxic agent in animals. Other evidence points to an interaction of metaldehyde and the γ-aminobutyric acid (GABA) (Sparks et al. 1996). Whilst the study focuses on vertebrate biochemistry, GABA has been shown to play important neurological roles in Mollusca including effects on motor, feeding and olfactory activities (Nezlin and Voronezhskaya 1997; Narusuye, Kinugawa, and Nagahama 2005; Moccia et al. 2009). The end result of metaldehyde interaction with the slug is the alteration of mucocyte cells, resulting in an increase in mucous production (Triebskorn 1989). The resulting dehydration causes loss of mobility, and eventual mortality after loss of 50% of body mass (Godan 1999).

Metaldehyde and methiocarbs are primarily formulated as edible baits, usually containing 2-8% toxicant; these are the familiar slug pellets used by both professional and amateur growers. The active chemicals are combined with wheat bran or barley flour to act as an attractant and feeding stimulant. The use of active substances at any higher concentration results in a progressive repellent effect, decreasing slug feeding. The bait attractant properties may be further enhanced via

addition of specific materials such as proteins (typically casein), or dextrose (Barker 2002). However, bait attractiveness is relative to other materials available for slug feeding; crops such as soft fruits are highly attractive to slugs, and are consequently more difficult to protect effectively. Additionally, the attractive effect of standard baits usually only lasts for around 3 days, depending on conditions, after which additional treatments are required. Since many slug species primarily live in the soil, and baits are applied on the surface, the effectiveness of slug baits is also compromised as protectants for crops which grow underground, such as potatoes and flower bulbs. In a given area, susceptibility to bait treatments for gastropod species can be less than 10% (Godan 1999). The limited effectiveness of bait treatments leads to very high rates of application, which in turn can have undesirable environmental consequences (see below; section 1.2.4).

Alternatives to bait treatments have been investigated, such as seed treatments. Seed treatments have the advantage of protecting plants during sprouting when they are most susceptible to damage, whilst using minimal active substance due to the small surface area of seeds. With damage to wheat most significant during seeding, metaldehyde seed treatments and sprays have been researched for potential in crop protection, but have shown poor results (Runham and Hunter 1970). More recent research into seed treatments of canola (*Brassica napus)* show that, in lab conditions, metaldehyde seed treatments are much more effective than equivalent baits (Simms, Mullins, and Wilson 2002). However, field conditions lead to a reduction in overall efficacy of treatments, primarily due to micro-organisms readily being able to utilise the metaldehyde as a carbon source. Whilst beneficial for long term ecological considerations, use of metaldehyde as a seed treatment requires the development of more complex formulations in order for them to be as effective (Simms et al. 2006).

1.2.3 Other molluscicides

A variety of chemical, biological and ecological alternatives are currently in use to protect crops from Molluscan species. As yet none have been shown to be effective enough that they can replace metaldehyde or methiocarb in an agricultural

setting. Carbamates chloethocarb and thiodcarb have been investigated for use with canola and winter cereals, though their properties are relatively similar to other carbamates such as methiocarb (Godan 1999). Iron phosphate-based chemical molluscicides can be considered the next best alternative to metaldehydes and methiocarbs. A number of metal salts were shown to have molluscicidal properties, and the research suggested that their efficacy could be improved by use of a chelation agent (Henderson and Martin 1990). Mini-plot trials indicate that for *D. reticulatum* and *Arion ater* iron phosphate can be as effective as metaldehydes. However other studies concluded that iron phosphate shows lower efficacy and higher cost than metaldehyde based crop protection methods (Speiser and Kistler 2002). Iron phosphate is the primary chemical molluscicide for organic farming, with many organic farmer associations completely prohibiting metaldehyde usage. The commercial molluscicide Sluggo® uses iron phosphate as an active ingredient and is marketed as organic and safe for pets and wildlife. Iron phosphate is combined with chelating agents such as EDTA to form iron chelates which are toxic to molluscs (Young and Armstrong 2001). Whilst less toxic to other organisms than metaldehyde, EDTA and iron phosphate baits have been shown to affect earthworm viability (Edwards et al. 2009) and cause toxic effects in mammals (Haldane and Davis 2009).

"Natural" molluscicides have also been described, such as garlic (allicin) (Schöder, Port, and Bennison 2004) and copper plate (Schöder, Port, and Bennison 2004), but their effects are limited to temporary repellent. Caffeine has been shown to cause mortality in molluscs, but in addition to being a psychoactive drug in humans and toxic to a wide range of non-target organisms, it has been shown to cause leaf damage to some crops when applied as a spray (Hollingsworth, Armstrong, and Campbell 2002). The nematode species *Phasmarhabditis hermaphrodita,* a known parasite of *D. reticulatum* was identified as a potential biological control. Field trials were conducted on asparagus (Ester, van Rozen, and Molendijk 2003) and since then the organism has been developed as a commercial product (Nemaslug®). But its application as a practical molluscicide is limited due to high price, short shelf life and temperature sensitivity (Speiser and Kistler 2002).

When compared with chemical molluscicides for mainstream crop protection the method has shown "poor results" (Rae, Robertson, and Wilson 2009).

1.2.4 Biological impact of molluscicides

Both metaldehyde and methiocarb are known to be toxic to all animals at high enough concentrations. There is a widespread belief that their use is hazardous to non-target organisms, and has a detrimental ecological effect, although evidence to support these claims is limited. Many invertebrate species have been shown to be unaffected by molluscicide bait poisons, though methiocarb is toxic to some carabids (Godan 1999). It is known that chemical baits are eaten by numerous other invertebrate species such as earthworms, beetles, centipedes and woodlouse (Gavin et al. 2007). Limited research is available on the effect of metaldehyde treatments on non-target invertebrates, such as long term effects on viability. The majority of metaldehyde poisoning of non-target organisms published comes from cases of toxicosis, and sometimes mortality, in various vertebrates species such as hedgehogs (Keymer, Gibson, and Reynolds 1991), cows (Valentine et al. 2007), dogs (Campbell 2008; Mills 2008), birds (Andreasen 1993) and humans (Shih et al. 2004; Bleakley et al. 2008). In most of these cases poisoning has occurred through consumption of very large amounts of slug pellets containing metaldehyde.

In many cases the risk of molluscicides to the environment is dependent on environmental factors such as frequency of treatments and weather conditions. Despite its break down in the soil by micro-organisms and relatively poor solubility, contamination of ground water is another issue. In some parts of the UK, water boards report metaldehyde levels at 10 times higher than 0.1 μg/L EU limit for water quality (Tao and Fletcher 2013). This level is not considered hazardous but water boards are failing to comply with European legislation and can face penalties for not maintaining water purity. Removal of metaldehyde from water is very difficult, adding further impetus for reducing its use.

1.3 Mollusca at the Molecular Level; Current Molluscan Genetic data

The reduction in cost and increased efficiency of next-generation sequencing has led to an explosion in the amount of genetic data available. This has included many mollusc species, though little is known about *D. reticulatum* save basic details such as having 30 chromosomes (haploid) (Fretter and Peake 1978). At the beginning of this project in 2009 there were less than 30 gene sequences known for *D. reticulatum*, with similar numbers seen for the majority of mollusc species, with a handful of larger EST sequence datasets for molluscs such as *A. californica*. Since then numerous parallel studies have taken advantage of high throughput sequencing on Molluscan species. There is now transcriptome data for a wide range of species including *Lymnaea stagnalis* (Feng et al. 2009), *Ruditapes philippinarum* (Milan et al. 2011), as well as large datasets for pearl oyster species (*Pinctada* sp.) (Huang et al. 2012). In addition to many transcriptome datasets, the first Molluscan genome was published for *Crassostrea gigas* (Zhang et al. 2012), with draft genomes for *Aplysia californica* and *Pinctada fucata* also available. Unfortunately there is no equivalent centralised repository such as flybase or wormbase for mollusc species. Efforts have been made to produce curated databases of multiple mollusc datasets such as MolluscDB (http://www.nematodes.org/NeglectedGenomes/MOLLUSCA/ [Accessed 26/02/13]). This dataset is primarily sequences determined by earlier technologies and does not include any recent datasets produced with next generation sequencing. The alternative is to use central databanks such as NCBI or EBI sequence databases. However surprisingly few of the sequence datasets for molluscs have available sequence data with annotation. The draft assembly of *Pinctada fucata* provides a genome browser (Takeuchi et al. 2012), but as yet the level of web and programmatic access to annotation for molluscs is still poor when compared with other invertebrate phyla, such as the flybase and wormbase projects.

1.3.1 Pyrosequencing

Pyrosequencing was developed in the mid-1990s as a sequencing technique initially utilised for SNP analysis due to its short read length (Rothberg and Leamon 2008). However with miniaturisation of components and increasing computer power

it is fast becoming the primary sequencing method for larger genetic studies. The high rate of improvements driven by competition between alternative technologies such as Roche 454, Illumina and Life Technologies' Ion-Torrent has driven down the price of sequencing. With the new wave of bench top machines and some technologies producing >600bp reads, this technology has replaced classic Sanger sequencing methods for many applications, and is bringing about a new revolution in genetic studies.

This project concerns the use of Roche 454 pyrosequencing technology. A summary of this process is as follows: Deoxyribonucleic Acid (DNA) is broken into small fragments and adaptors are ligated to each end. The fragments are ligated to small beads in an oil-water emulsion and amplified with polymerase chain reaction (PCR). The beads along with polymerases, luciferase and ATP sulfurylase enzymes are added to individual picolitre wells on a fibre optic slide (Margulies et al. 2005). Successive nucleotide triphosphates are added and washed out, A then C then G and T repeatedly; when the next nucleotide in the sequence attached to the bead is complementary to the nucleotide being washed, the strand is polymerised. This polymerisation releases inorganic phosphates which are converted to ATP by ATP sulfurylase, the ATP released is utilised by luciferase to convert luciferin into oxyluciferin with the reaction emitting photons. dATPαS rather than dATP is used for the polymerisation reaction as it is not a substrate of luciferase and will not interfere with the luciferin catalysis. Apyrase is used to degrade unincorporated nucleotides and ATP between each nucleotide wash (Ronaghi et al. 1996). Each well is monitored for photon emission and emissions are matched with the release of a specific nucleotide wash. Nucleotide repeats are detected via larger emissions resulting from multiple nucleotide inclusions in one wash (AA, AAA etc.). The adaptor ligated to each template sequence being polymerised is of known sequence and length and so can be used to normalise results for the rest of the sequence in the well. The sequence of flashes corresponding to nucleotide washes can be converted to a sequence with quality score, equivalent to a Sanger-style chromatogram generated per well.

Illumina's sequencing technology is a distinctly different sequencing process. Single stranded DNA fragments are generated from genomic or cDNA samples and 5' and 3' adaptors ligated to each end. The resulting fragments are randomly bound to the surface of a flow cell channel. A solid phase bridge amplification binds the unbound end of the DNA fragment to the flow channel surface. This creates a double stranded, bridged, fragment, which is then denatured into 2 single stranded DNA fragments 1 sense and 1 anti-sense, each attached at one end to the flow cell. This process can be repeated many times to create dense clusters of DNA all of the same sequence. The sequencing then begins by polymerisation, using nucleotides attached to coloured dyes, each nucleotide containing a different colour. The dyes block further polymerisation so that only a single nucleotide is incorporated. The dyes fluoresce with laser excitation and an image is taken of the flow cell. The resulting clusters can be identified as a sequence and the fluoresced colour representing the subsequent base for the sequence. The dyes are cleaved and the process is repeated to a sequence of approximately 75bp in length. The presence of both sense and anti-sense sequences allows for 5' and 3' sequencing to be done, allowing for up to 2x75bp in length to be sequenced; generating 'paired-end' reads.

Ion Torrent Semiconductor Sequencing shares a very similar method as pyrosequencing such as 454 sequencing. However the nucleotides incorporated during polymerisation are standard dNTP and after each successive wash of nucleotides, polymerisation is detected through release of hydrogen ions using an ion-sensitive field-effect transistor. This sequencing technology was released in February 2010 with a claimed read length of 50bp-100bp and was not available during our consideration of high throughput sequencing technologies for this project.

1.3.2 Sequence Data Assembly

De novo sequence assembly has been described as an unsolved problem. Unlike assemblies to reference data, *de novo* assembly relies upon overlapping regions of sequence within the dataset in order to align sections of nucleic acid (singletons) together into a contiguous sequence (contig). This has a number of disadvantages, including the reliance on having enough data to resolve overlaps as

well sections large enough to span tandem repeats. However high-throughput sequencing has enabled large enough quantities of data to allow for the assembly of many complete sequences. Additionally, advances in technologies have allowed much larger singletons, such as Roche 454 technology which has an upper limit of ~600bp. Whilst improvements of technology are generating better sequencing data; effective assembly software is also needed to assemble the data accurately.

There is currently a wide variety of assembly software available, including commercial, freeware and open-source programs. Deciding which assembler to use is not straightforward and is dependent on many factors such as type of data, size, and analysis requirements. A comparison using CAP3, MIRA, Newbler, SeqMan and CLC programs with Roche 454 data seemed to indicate Newbler or SeqMan as the best assembler (Kumar and Blaxter 2010). This analysis also suggested some benefit to merging contigs from multiple assemblies by a further assembly step, though no further benefit was seen by adding more assemblies. Alternatively some transcriptomes have been assembled via sequential, rather than merged, assemblies. This is where contigs and singletons from one assembler are used as the input in another assembly software, such as with the olive transcriptome (Muñoz-Mérida et al. 2013). This is relatively popular as is guaranteed to increase the overall size distribution of contigs compared to a single assembly method.

One problem with maximising contig length in assemblies, identified with a comparison of CAP3, MIRA, Newbler, and Oases on Roche 454 data, is the increase in the number of chimeric contigs (i.e. contigs containing two or more separate sequences; Mundry et al. 2012). This identified the problem that larger contigs are not necessarily better; an assembler could merely concatenate successive reads together and would score very highly for size distribution metrics. This analysis indicated that whilst Newbler restored the most full-length transcripts, it also produced the most chimeric contigs (artificially merged multiple transcripts together). Over-assembly of sequence data is likely to be further amplified by using sequential assembly throughput methods.

Newbler is in general considered the gold standard for Roche 454 sequence assembly (Ren et al. 2012), but flaws in its assembly have been highlighted. The

effectiveness of an assembly is relative to the metrics that are prioritised, and may differ depending on the desired outcome for the data. Modern assembly software is able to assemble sequences within a relatively short time frame on modest computing hardware, at least for smaller datasets, such as described here within. In conclusion, comparison between multiple software is likely advantageous, in order to choose from the selection the transcripts that seem closest to original sequences. For transcripts, this can often be determined by examining the predicted protein sequences.

1.3.3 Transcriptomics and Identifying Biochemistry

Transcriptome sequencing projects for non-model organisms are popular as they are substantially more efficient that genome sequencing projects. Unlike genomes, transcriptomes provide smaller datasets which cost less to sequence and are more computationally manageable. Additionally, with the sequencing of cDNA, the overall probability of sequencing coding regions is much greater. A large amount of the genome of an animal, whilst not devoid of information, containing for example r/tRNA, introns, promoters, enhancers, transposons, telomeres; has limited application when researching a non-model organism as compared to transcripts.

Previous studies have shown the benefit of using high-throughput sequencing techniques on crop pest digestive/gut tissues (Pauchet et al. 2009; Peng et al. 2011). These studies used homology searching to make functional predictions of for transcript sequences. The study sequencing the brown planthopper demonstrated the benefit of high-throughput sequencing by ordering contigs using number of reads attached to IPR terms (Peng et al. 2011). In this way the study identified chymotrypsins and trypsins as some of the most frequently sequenced proteases. The study demonstrated how large scale sequencing can be used to identify and focus upon candidate digestive enzymes, such as chymotrypsins, and thus predict key enzymes for the organism's biochemistry. Proteases can be subdivided into 4 mechanistic classes: serine, cysteine, aspartic or metallo-proteases, each of which can be affected by different types of protease inhibitors. In many cases an organism will

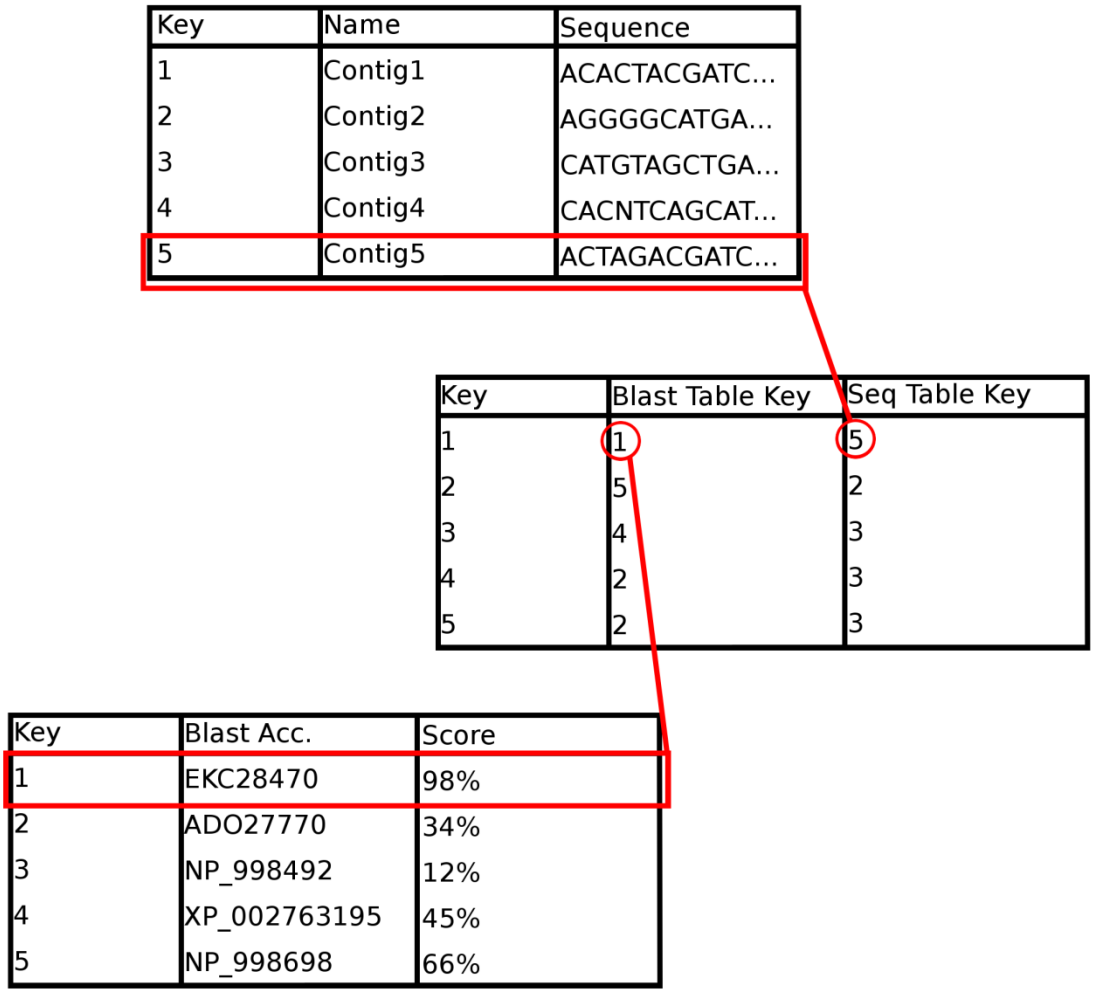
rely on one class of these proteases, knowledge of which can be used to as targets for crop defence strategies utilising inhibitors.

D. reticulatum has less than 30 nucleotide sequences in the NCBI/EBI databases, and only 3 independent protein sequences. Whilst some knowledge has been gathered to the likely importance of certain genes, no sequence information is available. With mechanisms such as RNA interference (RNAi) investigated as a crop defence mechanism (Price and Gatehouse 2008), knowledge of essential genes and protein sequences is necessary. Sequencing *D. reticulatum* transcripts to produce predicted protein sequences has a two-fold advantage. It allows the fast-tracking of experiments, bypassing the need to clone and sequence genes, but also allows the sequences obtained to be used to identify and focus in on the organism's biochemistry. With this data the predicted proteins from transcripts can be used to identify targets for protein inhibition and RNAi studies, or other methods to potentially create novel pesticides against the target pest.

1.3.4 Biological Databases and BioSQL

In many cases biological nucleic acid sequence data sets can be dealt with purely through the use of software available, which can manipulate data such as similarities determined by sequence comparison (BLAST) or gene ontology (GO) term files to produce overall statistics. However the limitations of these methods can be that the only statistics that can be generated are based on the individual functionality of the software in use. In many cases this leads to data being analysed based on what analyses can be done by software packages rather than what information a researcher wants. However, there is a growing amount of tools for scientific data management available to biologists to allow them to build analyses more tailored to their specific requirements (Katayama et al. 2013).

Outside of science the vast majority of institutes and companies manage large amounts of data using relational databases, which more often than not allow searching based on Standard Query Language (SQL). A relational database is a collection of tables of data types, tables can be considered similar to sheets in a spreadsheet. However unlike spreadsheets each table can be linked to other tables



Key	Name	Sequence
1	Contig1	ACACTACGATC...
2	Contig2	AGGGGCATGA...
3	Contig3	CATGTAGCTGA...
4	Contig4	CACNTCAGCAT...
5	Contig5	ACTAGACGATC...

Key	Blast Table Key	Seq Table Key
1	1	5
2	5	2
3	4	3
4	2	3
5	2	3

Key	Blast Acc.	Score
1	EKC28470	98%
2	ADO27770	34%
3	NP_998492	12%
4	XP_002763195	45%
5	NP_998698	66%

Figure 3 : Simple example of a relational database model

The 3 tables here represent a simplified model of a relational database. In this example the top and bottom tables contain sequence and BLAST annotation data. The sequences are linked to the BLAST annotation via a third table which holds both the key for the sequence and the BLAST annotation in a row. In this case a sequence can be linked to multiple BLAST annotations and equally a single BLAST annotation can be linked to multiple sequences. In this case we can select all sequences that are linked to BLAST accession AD027770 by retrieving the key for the accession and selecting all rows in the middle table with the key 2 in the BLAST table column and the returned rows will then contain all the keys for all sequences attached to this BLAST annotation. Many more tables and columns can be added and the data retrieved based on a specific column of data such as BLAST annotation score.

using a key, almost always numerical, see Figure 3. By linking rows from different tables together via keys a database can represent a multi-dimensional set of data in a way a flat spreadsheet cannot. This can be particularly useful for biological data, with many biological data such as proteomes and reactomes also having this type of data structure. Other data structures such as taxonomies which represent hierarchical data structures can still be represented in relational databases using models such as Nested Set, Adjacency List or Path Enumeration (Celko 2004). Additionally the inclusion of SQL in most relational database management software (RDMS) allows a powerful means of querying linked tables and filtering rows and columns for specific data.

There are a number of biological database schema (blueprints for database structures) now available for managing biological data such as the Chado (Mungall, Emmert, and The FlyBase Consortium 2007) or BioSQL (Katayama et al. 2010) schema. Flybase is an example of a well-known biological database which is built on top of a relational database using SQL and provides both web and programmatic access to a wide range of different sequence information and annotation. Whilst these schema are not necessarily optimal for managing all types of biological data, as bioinformatics blogger Brad Chapman has identified, these databases can always be modified to better suit biologists' needs (Chapman 2013).

An example of the basic SQL syntax for selecting a column in a table would be “SELECT column FROM table” where table and column were replaced with specific identifying names. This quickly becomes very useful with a query such as “SELECT sequences FROM table WHERE length > 100” which filters out short sequences. Producing a list of all sequences >100bp in a dataset of many thousands of sequences becomes trivial, once the database is set up. In general SQL queries used in this project are less human readable but can still be understood, one of the most complex used in the project is the following:

```
SET @runtot:=0; SELECT q1.EVALUE, (@runtot := @runtot + COUNT) AS CUMULATIVE FROM (SELECT bioentry_dbxref.evalue AS EVALUE, COUNT(bioentry_dbxref.evalue) AS COUNT FROM bioentry_dbxref WHERE bioentry_dbxref.run_id=x AND hit_no=1 AND bioentry_dbxref.rank=1
```

```
GROUP BY bioentry_dbxref.evalue ORDER BY bioentry_dbxref.evalue
ASC;) AS q1
```

This query can be written in English as: Create a running total, get the expect value and count that expect value, and increment the running total with that count. Only get records which are part of the relevant BLAST run and are the top BLAST hit and top hit region, order the data so that the expect value is ascending in order.

The output of this data is a two column table which contains in the first column, the expect values and the second, the running count of BLAST annotations, as shown in Table 1. This data can then be used in a basic x/y graph and represents the distribution of BLAST data below a given expect value. (100% of hits are below an expect value of 10, representing random similarity, and 0% hits are below the expect value representing complete sequence identity where expect ≈ 0). Using these types of queries statistics about the entire dataset can be generated nearly instantly, with queries taking only a few seconds to run across millions of rows of data.

RDMS is not a replacement for software such as BLAST sequence search utilities, which uses Berkeley DB database and is not a relational database. The

Expect value	**Cumulative Count**
1e-99	1
1e-89	3
1e-60	4
...	...
10	1121

Table 1 : Example of output from an SQL query on the BioSQL database

implementation of this database model in this project is for the same application as its use in flybase, which is to pull together many different forms of data generated by programs like BLAST. The data is uploaded by first parsing files such as XML

output files produced by BLAST and INTERPRO done by programmatically reading the file line by line and pulling out relevant information. This is then uploaded by using SQL INSERT statements which push data from the file parser into the databases. The implementation of this can be variable and there are numerous templates in biological software packages such as bioJava and bioPerl which can be modified to suite the specific requirements of the data input script. Whilst the upload of data can be time consuming and a variety of checks need to be done in order to verify the data has been uploaded correctly. Once the database is established it provides a mechanism for extracting any annotation data generated using a keyword or specific data filter. In addition cross-references with other databases are much easier to perform and results produced rapidly.

1.4 RNA Interference and Crop Protection against Molluscs

RNA interference (RNAi), the gene silencing mechanism was first discovered in petunia plants in 1990. The effect was more fully understood later in the nematode *Caenorhabditis elegans* in the late 1990s where dsRNA rather than ssRNA was identified as being responsible for gene silencing (Sen and Blau 2006). The method has since become a powerful tool for down-regulating target genes. The effect was shown when double stranded RNA (dsRNA) containing a strand complimentary to a mRNA was injected or fed to *C. elegans* leading to down regulation of the corresponding gene (Fire et al. 1998). The effect can also be produced by "anti-sense" RNA, a single strand RNA complementary to mRNA, which results in dsRNA being formed. RNAi has been shown as an endogenous pathway ubiquitous across most eukaryotic organisms; a general overview is shown in Figure 4.

The essential active component of the RNA interference pathway is the double stranded small (23bp) siRNA fragments without which no down-regulation can occur. siRNA provides the guiding strand of nucleic acid which is used to target specific gene products, stage (C) in Figure 4. siRNA can be introduced into cells to stimulate interference, but for invertebrate studies longer dsRNA, which is more stable and easier to work with, can be used. For some nematodes (including *C. elegans*) dsRNA can be transferred to the organism simply via introduction to the

surrounding media. Alternatively various methods such as injection, transfection or electroporation can be used for initial uptake into the cellular space of the organism (May and Plasterk 2005). Both *C. elegans* and insects such as the fruit fly *Drosophila melanogaster* have been shown to uptake dsRNA into cells after its introduction to the extracellular space (stage (A) in Figure 4). This is then cut into siRNA duplexes of around 21-27 bp by the Dicer RNaseIII-type enzyme in combination with a RNA binding proteins, such as Loquacious in *D. melanogaster*, stage (B) in Figure 4.

Whilst higher organisms also contain Dicer functionality immunological responses to longer dsRNA fragments can interfere with results. In mammals large dsRNA fragments (>30bp) cannot be used as they evoke an interferon response or a non-specific inhibition of protein synthesis through dsRNA-dependent protein kinases in mammals (Buckingham et al. 2004). Studies also demonstrate interferon like responses in some other non-mammal vertebrates, leading most studies to use siRNA (Sifuentes-Romero, Milton, and García-Gasca 2011). Use of dsRNA has been more successful in nematodes and insects, where no equivalent immunological response to large dsRNA fragments is seen.

After either introduction of siRNA or long dsRNA and processing to siRNA, siRNA is incorporated into the RNA-inducing silencing complex (RISC). The antisense strand of the siRNA is used as a guide strand to bind to complimentary mRNA. An argonaute protein forms the primary catalytic protein in the RISC and endonucleolytically cleaves the target RNA between the 10th and 11th base relative to the guiding strand (Martinez et al. 2002). Systemic effects, where the RNAi effect is transmitted to other cells, have also been shown in *C. elegans* and some plants. However, in *D. melanogaster* no systemic effect has been shown likely due to the lack of a RNA-directed RNA Polymerase (RdRP) homologue, which is necessary for persistence in *C. elegans* (Sijen et al. 2001), through amplification of the siRNAs. RdRP homologues have not been found in any insect genomes sequenced, but evidence for systematic effects in other insects has been presented (Tomoyasu et al. 2008). Discussion in the area continues with recent studies identifying the difficulty

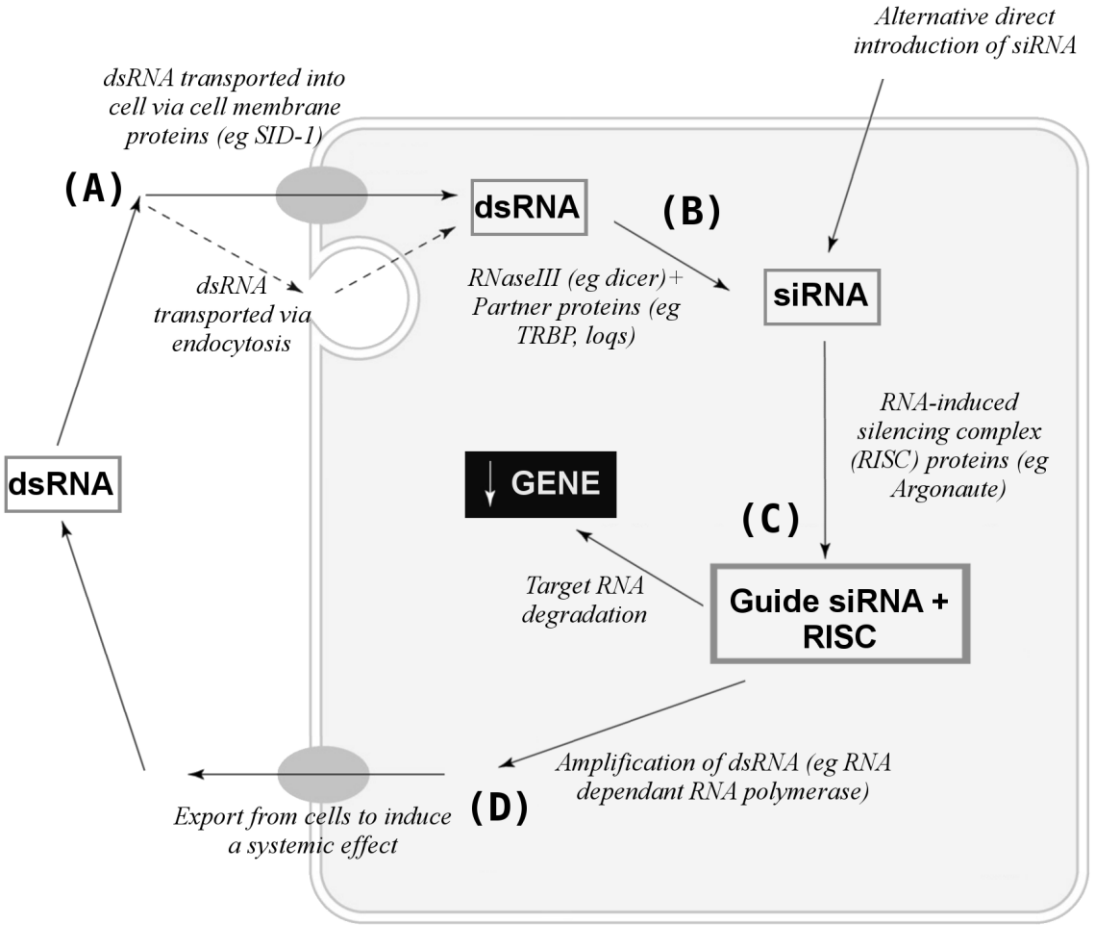


Figure 4 : General stages of RNA interference

(A) dsRNA is initially introduced into the extracellular space. It is then transported into the cell via endocytosis and/or cell membrane channel mediated transport such as in C. elegans, via the SID-1 protein. (B) dsRNA is cut into small dsRNA duplexes (siRNA), the size of which depends on the distance between the nuclease catalytic site and PAZ domain in the dicer protein. siRNA can also be directly introduced to the cell, as is done for organisms with immunological reactions to large dsRNA fragments, such as Humans. (C) RNA-induced silencing complex forms with the siRNA, the none-guide strand is discarded and the complex then goes on to stimulate degradation of transcripts complimentary to the guide siRNAs. (D) In organisms such as C. elegans, the dsRNA is also amplified and exported out of the cell, through channels proteins like SID-1 and elicits a systemic effect.

in distinguishing systemic effects from a very efficient system of cellular uptake and storage of dsRNA (Miller et al. 2012).

Even without systemic effects, RNAi effects have the potential to be a useful method for causing changes in physiology specific to target organisms. Gene knock-down, the reduction in levels of mRNA for specific genes and consequently the protein products, can lead to a variety of physiological effects, such as developmental arrest, failures of gut physiology, detoxification biochemistry, or direct mortality. In the field of crop protection, pests can be targeted by through expression of dsRNAs in transgenic plants, which is emerging as a crop protection technology of wide applicability. There are a variety of obstacles to consider with the use of dsRNA, the primary one being introduction into the organism. Host Delivered RNAi (HD-RNAi) has been suggested as a solution, and shown some success (Fairbairn et al. 2007). It also bypasses the problem, which many *de novo* crop protection methods have, of degradation over time, as dsRNA can be continually produced. The first stage in distinguishing RNAi against Mollusca would be to design and prove dsRNA targets against the organism. The first stage is injection, in order to actually ascertain if a knock-down effect can be elicited, followed by feeding assays. The long-term goal of these experiments would be to generate working targets which could be used in a system such as HD-RNAi.

1.4.1 RNAi against Mollusca

Genetic data for a number of molluscs indicates the presence of RNA interference machinery. However *C. elegans*-like dsRNA based silencing, is less clear, as much of the Molluscan genetic data is fragmented with limited annotation. Dicer is one of the most well conserved proteins with its mode of action necessitating a number of different domains. This provides a relatively unique signature to look for with PAZ, dsRNA binding and dsRNA-specific endonuclease domains required and conserved between *C. elegans*, *D. melanogaster* and *H. sapiens* (Lau et al. 2012). Transcript fragments from *L. stagnalis*, *A. californica* contain Paz domains, which are found in argonaute (AGO) and dicer protein families (Cerutti, Mian, and

Bateman 2000). One complete dicer protein comes from *C. gigas* and includes all the expected protein domains (Genbank: EKC26346).

The recent sequencing and annotation of the *C. gigas* genome has provided a complete set of genes to search through. Homologues for RNA-dependent RNA polymerase (Genbank: JH817893) and SID-1 transmembrane dsRNA transport protein (Genbank: EKC42950) are available in the *C. gigas* genome. However, despite the presence of at least some of the machinery homologous to that which powers the systemic RNAi effect seen in *C. elegans*, there are very few examples of the RNAi effect in molluscs, reflecting the fact that evidential RNAi studies in any non-nematode/insect invertebrate are relatively scarce. In most cases there are only a handful of studies in for each phyla, primarily demonstrating the presence of an RNAi effect. Evidence from the variability of insect RNAi results, where RNAi effects differ from order to order, or even from species to species, means using RNAi data from one mollusc species as evidence for its effect in another is problematic.

The small number of papers that are available for the Molluscan phylum does provide evidence for RNAi effects in these organisms. The earliest publication appears in from 2001 and initially proves RNAi effect by co-micro-injection of a DNA expression vector containing the Luciferase gene and a dsRNA fragment complimentary to the gene into *Aplysia californica* giant neurons (Lee et al. 2001). Transient expression of luciferase, visualised by light emission, which was observed when vector alone was injected, was down-regulated by the dsRNA. This provides evidence that once dsRNA is within the cell, a gene knock down effect can be elicited, although micro-injection bypasses the need for uptake of the dsRNA from the intra-cellular space into the cell.

Experiments which didn't involve micro-injection of dsRNA into cells also indicate that RNA interference occurs; most workers have attempted to inject dsRNA into the body cavity of molluscs, but a variety of conditions have been used, which means that the results are not fully comparable from experiment to experiment. Even though direct comparisons are not possible for all these experiments using injected dsRNA, it is apparent that the observed RNAi effect is very variable between mollusc species, and between genes targeted. There is also disagreement between

different experimenters over the timescale of RNAi effects. One study shows a phenotypic effect within 3 hours post injection (Korneev et al. 2002), but most of the quantitative PCR based methods shows a measurable effect on transcript levels only after 24 h. RNAi experiments in *Chlamys farreri* and *Biomphalaria glabrata* show an effect beginning at 24h post-injection, with the maximum knock down in gene expression before 96 h (Wang et al., 2011; Jiang et al., 2006). The study on *B. glabrata* continued to monitor gene expression levels for *FREP2* and myoglobin genes until complete cessation of any RNAi effect by 10 days post injection. This contrasts with a study on *C. gigas*, where the first time point was at 9 days, with their data suggesting an RNAi effect was still maintained up to a month post-injection (Fabioux et al. 2009). This result showing long-tern RNAi effects suggests a potential systemic effect for *C. gigas,* and would agree with the genomic data previously discussed. The combination of lack of data, and differing results from data available, means effectiveness of RNAi against *D. reticulatum* is not easily predictable.

1.5 Aims and Objectives of the Project

Investigation

1. Identification of the best methods for isolation of good quality RNA for the production of cDNA from *D. reticulatum* tissues.
2. Production of cDNA from digestive gland and random sequencing of cDNA in order to verify it as digestive gland transcripts.
3. Use of RACE as a proof of principle procedure to demonstrate amplification of full genes from PCR product fragments produced from redundant primer PCR.
4. Use of redundant primer PCR to isolate ion channel gene sequences.
5. Conduct high throughput sequencing of digestive tissues and neural tissue and assemble into a workable database in order to extract sequence information and relevant annotation.

Analysis

1. Process high throughput sequence data through previously described annotation pipelines
2. Assess the digestive gland for presence of digestion related proteins through functional prediction based on homology.
3. Assess the neural tissue dataset for neural related proteins as done with digestive dataset.
4. Assess the functionally predicted proteins for potential further research and development.

Application

1. Utilise previous analysis of sequence to identify potential targets for RNAi
2. Clone vector-insert RNAi expression constructs for dsRNA production.
3. Synthesise dsRNA and assay it against *D. reticulatum*.
4. Investigate other methods for producing a molluscicidal effect such as production of a protein substrate to assay against the target organism

Chapter 2 | Materials & Methods

2.1 General molecular biology methods

All chemicals and reagents supplied by Sigma (St. Louis, USA) or VWR (BDH) (Poole, Dorset UK). Unless otherwise stated solvent was autoclaved/filter sterilised deionised water or distilled water. Where protocols are not elaborated they can be consider standard across most biology laboratories.

2.1.1 Recipes

Bacterial culture media	LSLB broth: 0.5% (w/v) NaCl, 1% (w/v) tryptone, 0.5% (w/v) yeast extract, prepared in distilled water. LSLB agar: 1.5% Bacto agar (Difco) added to LSLB broth
Agarose gel electrophoresis:	TAE (50X): 2 M Tris/Acetic acid pH 7.7, 50 mM EDTA
DNA loading buffer:	10 mM Tris/HCl pH 8.0, 10 mM EDTA, 30% (w/v) glycerol, 0.1% (v/v) Fast Orange G, prepared in distilled water
Protein gel electrophoresis (SDS-PAGE)	5X SDS sample buffer: 0.5 M Tris/HCl (pH 6.8), 50% (v/v) glycerol, 5% (w/v) SDS, 0.005% (w/v) bromophenol blue
Acrylamide	30% (w/v) acrylamide: 0.8% (w/v) bis-acrylamide stock solution (37.5:1) (Protogel, National diagnostics)
Resolving buffer	3.0 M Tris/HCl pH 8.8
Stacking buffer	0.5 M Tris/HCl pH 6.8
Reservoir buffer (10x)	0.25 M Tris/HCl pH 8.3, 1.92 M Glycine, 1% (w/v) SDS
(CBB) Stain	40% (v/v) Methanol, 7% (v/v) glacial acetic acid,

	0.05% (w/v) Coomassie Brilliant Blue (CBB)
Destain	40% (v/v) Methanol, 7% (v/v) glacial acetic acid
Stacking gel mixture	2.5% Protogel (37.5 : 1 acrylamide : bisacrylamide; National Diagnostics), 125 mM Tris-HCl (pH 6.8), 0.1% (w/v) SDS, 0.1% (w/v) ammonium persulphate, 0.0075% (v/v) N, N, N', N'-tetramethylethylenediamine (TEMED)
Resolving gel mixture	(12.5% or 15% or 17.5 % (w/v) Protogel, 375 mM Tris-HCl (pH 8.8), 0.1% (w/v) SDS, 0.075% (w/v) ammonium persulphate, 0.05% (v/v) N, N, N', N'-teretramethylethylenediamine (TEMED)
Protein molecular weight marker – KiloDalton (kDa)	SDS7[A] (Sigma): 66 kDa Bovine albumin, 45 kDa Egg albumin, 36 kDa Glyceraldehyde- 3-phosphate, 29 kDa Carbonic anhydrase bovine erythrocytes, 24 kDa PMSF-treated trypsinogen, 20 kDa Soybean trypsin inhibitor, 14 kDa α-lactalbumin SDS7 [B] Pierce Unstained Protein Molecular Weight Marker.
Western blotting Bjerrum & Schafer-Neilson buffer	48 mM Tris-HCl, 39 mM Glycine, 20% (v/v) methanol, 0.0375% SDS, pH 9.2
Ponceau stain	0.1% Ponceau S, 5% acetic acid in distilled water
Phosphate Buffered Solution (PBS) 10x	0.015 M KH_2PO_4, 0.08 M Na_2HPO_4, 1.37 M NaCl in distilled water
Blocking solution	5% Non-fat milk powder, 1X PBS, 0.1% Tween-20
Anti-Sera solution	5% Non-fat milk powder, 1X PBS, 0.1% Tween-20
PBST	1X PBS, 0.1% Tween-20
Chemiluminescent	Solution A: 100 mM Tris/HCl pH 8.0, 0.2 mM

detection reagents (Solution-A)	coumaric acid, 1.25 mM luminol in 50 ml distilled water
Chemiluminescent detection reagents (Solution-B)	10% H_2O_2 (30% solution) in distilled water
DNA molecular weight marker	Lambda DNA digested with Eco471 (AvaIII) or HyperLadder I (Bioline)
10x PCR buffer	400mM Tricine-KOH (pH 8.7), 150mM KOAc, 35mM $Mg(OAc)_2$, 37.5µg/ml BSA, 0.05% Tween 20, 0.05% Nonident-P40
DEPC Water	Diethylpyrocarbonate (DEPC) was added to water from Milli-Q (Merk) water purification system to a final concentration of 0.1% (v/v) and left at 37°C for 2 hours. Water was then autoclaved for 20 minutes.

2.1.2 Phenol-chloroform ethanol precipitation

Where phenol-chloroform ethanol precipitated is stated, steps 1-3 are done, where chloroform extracted is stated only step 2 and where ethanol precipitated is stated, only step 3 is done.

1) Nucleic acid solution was first vortexed with Phenol:Choroform:Isoamyl Alcohol (25:24:1) and centrifuged for 15 minutes at 14,000 rpm. Aqueous layer was transferred to a new tube, hydrocarbon layer was discarded.

2) Solution was treated as in step 1 but with 100% chloroform.

3) 10% v/v of 3M Sodium acetate pH 5.2 and 250% v/v absolute ethanol were added to the nucleic acid solution for precipitation and incubated overnight at -20°C. Solution was centrifuged at 12,000 g for 15 minutes at 4°C. The pellet was washed with at least 5 times original volume of 70% ethanol twice. The nucleic acid pellet was air dried and re-suspended in appropriate amount of nuclease free water,

sometimes dissolution was done by incubating nucleic acids at 37°C for 15-30 minutes.

2.1.3 Oligonucleotides

Oligonucleotides were synthesised by Sigma Genosys Service and used as PCR Primers. Primers were resuspended in sterile distilled water to a concentration of 100 pmol/µl and stored at -20°C. In standard PCR reactions the primers were used at final concentration of 0.2 to 1.0 µM depending on primer types.

Melting temperature for the primers was calculated by the following formula T_m was calculated using the OligoCalc online tool (Kibbe 2007). Primers were used at 3°C below the calculated T_m for Taq and Taq based polymerase and 4°C above for Phusion HF polymerase [Fermentas].

2.1.3 Degenerate Primer Design

Degenerate primers refer to primer mixes containing multiple possible similar sequences with minor variations at certain base positions. Degeneracy represents the number of possible combinations of base pairs and can be calculated as the product of each individual base's degeneracy where A,T,C,G represent the minimum degeneracy of 1 and N (any nucleotide) a maximum of 4. Primers were designed to have a minimal amount of degeneracy in order to increase the specificity of the PCR amplification. Whenever possible, to allow efficient primer extension, oligo dC or dG residues were preferentially incorporated at the 3' end of the primer. Due to the degenerate nature of the primers it was necessary to increase primer concentration in PCR reactions, so primers matching the target sequence were present at a sufficient concentration to allow amplification. Degenerate primers were added to PCR reactions at a concentration of 1.0 µM, instead of the standard 0.2 µM.

2.1.2 DNA amplification with Polymerase Chain Reaction (PCR)

Standard PCR reactions were conducting using 0.2 ml & 0.5 ml tubes performed on a Perkin Elmer 2400 thermal cycler. Taq PCR reactions were conducted in 25 µl or 50 µl reactions with 10x PCR buffer (see recipe), 0.2 mM dNTPs, nucleic acid template (~0.1-2 ng/µl), 0.2 µM DNA nucleotides, and 1.25 U

Taq polymerase. For Taq PCR, the standard PCR program was Step1: 94ºC at 1min; Step2: 25 cycles { 94ºC at 30sec, Tm-3ºC at 30sec, 72ºC at 30sec/kb+30sec }; Step3: final 72ºC at 7mins; were used unless otherwise stated. For amplification of sequences for fragments for RNAi constructs Phusion Polymerase was utilised, for all other PCR amplification Taq polymerase was used.

2.1.3 High Fidelity PCR

Where 'high-fidelity' PCR is referred to, these PCR reactions utilised proof reading enzymes, the reactions used High-Fidelity Phusion Polymerase [Fermentas] with PCR reactions setup following manufacturer's instructions.

2.1.4 Touch Down PCR Protocol for degenerate primers

Touch-Down PCR follows the same protocol as standard PCR except the cycling steps are altered so that the annealing temperature decreases each cycle. Touch-Down priming has the advantage of reducing spurious priming (Don et al. 1991) which is a particular problem for degenerate primers. For the successful degenerate primers used for amplifying the *D. reticulatum*, the successful cycling protocol was as follows. Step1 94ºC at 3min; Step2: 20 cycles { 94ºC at 30sec, 50ºC-0.5.cycle^{-1} at 30sec, 72ºC at 2min }; Step3: 13 cycles { 94ºC at 30sec, 40ºC at 30sec, 72ºC at 2min }; Step4: final 72ºC at 7mins.

2.1.5 Colony PCR

Colony PCR reactions were conducted to screen colonies transformed with plasmid DNA constructs for the presence of a PCR insert within the plasmid cloning site. The volumes were reduced to 25 µl with components scaled accordingly, usually made as a master mix. A PCR product of the insert, previously validated, was used as a positive control and empty plasmid DNA vector as a negative control. Colonies were picked from an agar plate containing a relevant antibiotic with plastic pipette tips and left in aliquots of PCR mix for 1-2 minutes. A 'master plate' was used to keep track of colonies, with the tips being smeared first on the 'master plate' before being put into the PCR mixture aliquots. The pipette tips were then removed from tubes and the PCR then conducted with the same cycling as the original PCR.

This allowed for transfer of cells from the pipette into the PCR mixture. These cells would lyse in the first cycling step and release plasmid DNA which can then be primed and polymerized if the priming sites exist within the plasmid sequence. In this way clones can be assayed for the presence of an insert directly without necessity for a plasmid extraction step which would require a further overnight incubation period. The resulting products of the PCR are then assessed by gel, with positives indicating a probable successful transformation, which can be further confirmed by sequencing the clone.

2.1.6 Nucleic Acid Quantification

Nucleic acids were purified with phenol-chloroform extraction and ethanol precipitation and resuspended in nuclease free water or relevant buffer. The nucleic acids were then quantified using a Thermo-Scientific NanoDrop™ 1000 Spectrophotometer under highly accurate UV/Vis analyses of 1 µl samples with nucleic acid free equivalent media which was used as a blank measurement. Wherever possible, nucleic acids were also quantified by comparison with nucleic acid of known sample concentration on gel electrophoresis with quantitation done using the ImageJ software package available at http://rsb.info.nih.gov/ij/ [Last Accessed 06/05/13].

2.1.6 DNA ligations

DNA ligation of vector and insert with compatible ends was done using T4 DNA Ligase and Ligase Buffer [Promega] following manufacturer's guidelines. Quantities of vector and insert DNA were calculated based on a 1:1 molar ratio which was estimated by relative size in base pairs multiplied by mass of DNA. Ligation mixtures were incubated at 4ºC overnight or room temperature for 1-2 hours unless otherwise stated.

2.1.7 Isolation of plasmid DNA

Bacterial cell colonies containing plasmids were picked from agar plates or glycerol stocks and used to inoculate 10ml LSLB with appropriate antibiotics in 100ml McCartney bottles. Cultures were incubated on shaker at 150rpm overnight at

37°C and then centrifuged at 4000 g for 15 minutes. The supernatant was discarded and plasmids were then extracted from cell pellets using the Wizard® *Plus* SV Minipreps DNA Purification Systems [Promega] following the manufacturer's guidelines. The resulting plasmids were resuspended in nuclease free water and stored at -20°C for further use.

2.1.8 Restriction endonuclease digestion of DNA

Restriction enzyme (RE) digestions were carried out using commercially available RE enzymes [Fermentas, Promega, NEB or Roche] and were carried out using buffers and temperatures recommended by the manufacturers. Typically digests were carried out at 37^{o}C on 1 µg of DNA using 2-10 units of RE (under optimal conditions 1 U of RE will completely digest 1 µg of DNA in a 50 µl reaction volume in 1 hour). Where double digestions were conducted optimal buffers were selected via manufacturer's guidelines. Wherever possible single digestions of each enzyme in buffer were conducted to provide diagnosis for failed double digestions. Restriction products were separated by Agarose gel electrophoresis and visualised by ethidium bromide staining.

2.1.9 Agarose Gel electrophoresis

Samples were loaded with DNA loading buffer (see recipe) in a 0.8% TAE Agarose gel, suspended in an electrophoresis tank filled with TAE buffer, with 0.05% ethidium bromide added to both gel and buffer. The samples are electrophoresed at 100V until the banding pattern has separated sufficiently, usually around 30-40 minutes. The protocol follows common laboratory practice (Sambrook and Russel 2001).

2.1.10 Gel Extraction of DNA

DNA Agarose gels were visualised on a trans-illuminator (UVB, λ 302 nm) and appropriate bands were excised from the gel by using a single edged razor. DNA was purified from the Agarose by using QIAquick gel extraction kit [Qiagen], according to manufacturer's instructions. Eluted DNAs were stored at –20ºC.

2.1.11 Agarose Gel electrophoresis for size separation of cDNA

cDNA samples, and marker, were loaded as with standard Agarose gel electrophoresis and gel electrophoresis conducted until cDNA is well distributed across the gel. The gel is taken out of the tank and dissected into two pieces based on predetermined marker size (2Kb). The low molecular weight half of the gel is discarded and the high molecular weight half is returned to gel tank and run with reversed polarity to return cDNA to a more compact form. Before the majority of the cDNA reaches the well the section of gel, the band is excised and gel extracted as described in a previous section.

2.1.12 SDS-PAGE electrophoresis

Protein samples along with molecular weight marker (SDS7) were denatured and reduced before loading by diluting in 1x sample loading buffer and boiling for 10 minutes. Denatured proteins were separated according to their size by sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) according to (Laemmli 1970). Mini-gels (9x10 cm) were run in 1x reservoir buffer at constant voltage (100-150 V) in ATTO-AE450 apparatus. Following electrophoresis, gels were either stained of transferred to nitrocellulose membrane for western blots.

2.1.13 Staining with Coomassie Brilliant Blue

After SDS-PAGE, proteins in the gel were visualized by staining with CBB stain for minimum 3 hours, followed by destaining with destain until the background is clear. Both staining and destaining were carried out at room temperature with gentle agitation.

2.1.14 Western Blotting

Proteins were transferred to nitrocellulose (Hybond ECL, Amersham) membranes followed by electrophoresis by electro-blotting, using a standard semi-dry transfer method. Gels to be blotted were equilibrated in Bjerrum and Schafer-Neilson buffer (see recipe) by soaking for 30 minutes at room temperature. Nitrocellulose membrane and 3MM blotting papers (Whattman) were cut to the same

dimension as gels and pre-wet in same buffer. The blot was set up on ATTO AE-6675 blotting apparatus. Electroblotting was conducted at constant current set to 125-150 mA (2.0 mA/cm^2) for 60 minutes. Efficiency of transfer was checked by staining membrane with Ponceau S stain, molecular weight marker bands were marked with pencil and the membrane was destained with distilled water.

2.1.15 Chemiluminescent detection of membranes

Non-specific protein binding sites on the membrane were blocked by incubating in 50 ml blocking solution for 1 hour at room temperature with 3 changes and gentle agitation. After blocking membranes were reacted with appropriate dilution (generally 1:3000 unless otherwise specified) of primary antibodies (Anti His) in antisera for 2 hours at room temperature or overnight at 4°C with gentle agitation. Unbound primary antibody was removed by washing with antisera for 30 minutes with three changes at room temperature. The membrane was transferred to antisera containing secondary antibody (Goat Anti Mouse IgG (H+L)-HRP conjugate) in appropriate dilution (generally 1:3000 unless otherwise specified) and incubated for 1-2 hours at room temperature with gentle agitation. Membrane was washed with 1x PBST for 30 minutes with three changes to remove unbound secondary antibody. Excess Tween 20 was removed by washing membrane briefly with distilled water. Enhanced Chemiluminescence (ECL) reagents [Amersham] were used to detect specifically bound secondary antibodies. Solution A (5 ml) was mixed with solution B (15 µl) shortly before exposure of membranes. Specific antibody binding was visualised by exposing membranes to photosensitive film (Fuji-RX). Exposed films were washed and developed with an automatic developer (X-ograph Imaging Systems Compact X4).

2.1.16 Glycerol stocks of *E. coli* strains

Single colonies of recombinant *E. coli* containing DNA plasmids were used to inoculate 10 ml LSLB cultures which were grown overnight at 37°C. The culture was centrifuged at 4000 g for 10 minutes at room temperature. Supernatant was removed, and cells resuspended in 800 µl solution LSLB 70% (v/v), Glycerol 20%

(v/v), Distilled Water 10% (v/v), vortexed briefly, transferred to 1ml screw-cap tubes and frozen at -80°C.

2.1.17 *E. coli* transformation

All *E. coli* strains, TOP10 [Invitrogen], Origami B(DE3) [Novagen] and HT115 (*rnc14::ΔTn10*), were prepared as electrocompetent cells as per standard laboratory protocol (Sambrook and Russel 2001). Unless otherwise stated all plasmids were transformed using 1 µl of either a ligation mixture or sufficiently diluted purified plasmid. All transformations of ligated plasmids in the ligation reagents were first chloroform-extracted, by vortexing with equal volume of chloroform and centrifuging at maximum speed on a bench centrifuge for 5 minutes. Aqueous layer was transferred and mixed with 50 µl competent cell aliquot and transformed at following conditions: electrical pulse set to 25 µF, capacitance set to 1.8 kV and 200 Ω resistance. Cells were transferred to 1 ml of LB broth and then incubated at 37°C for 1 hour, with shaking (220 rpm). Transformants were selected by plating cells (5-10% of volume) on LB-agar containing an appropriate antibiotic.

2.2 RNA and cDNA

All RNA and cDNA experiments were performed with either nuclease free water or DEPC treated water. Where possible equipment used in RNA extractions was washed with DEPC treated water and oven baked at 200°C.

2.2.1 Preparation of digestive gland total RNA for SMART cDNA

Total RNA used for SMART cDNA synthesis kit [Clontech Laboratories, Inc.] was produced using RNeasy Mini Kit [Qiagen]. Tissue was extracted from a single adult slug by dissection of the digestive gland which coils around the *D. reticulatum* intestinal tract; care was taken to remove any non-digestive tissue associated with the gland. Tissue took approximately 5 minutes to dissect from organism and transfer to next stage. Approximately 60 mg of tissue was transferred to the first stage solution of the RNeasy kit [Qiagen] and homogenised using a pestle drill piece and variable speed laboratory motor [TRI-R Instruments, Inc.]. The RNA extraction continued following the manufacturer's guidelines. The resulting RNA

was quantified with a NanoDrop™ 1000 Spectrophotometer [Thermo-Scientific] and transferred immediately to the first strand synthesis step of SMART cDNA synthesis kit. It was found that the time between the dissection and first strand synthesis was paramount to good quality cDNA, as such all efforts were taken to conduct experiments as quickly as possible.

2.2.2 Preparation of neuronal tissue for SMART cDNA

Total RNA used for SMART cDNA synthesis kit [Clontech Laboratories, Inc.] was produced using RNeasy Mini Kit [Qiagen]. The tissue dissected was primarily the nerve ring, situated around the crop of the organism and any attached nerves that could be dissected along with the ring. Due to the size of nerve ring, to increase the total weight, optic sensory tentacles, which attach directly to the nerve ring ganglion bundle were also dissected and included, as they contained large nerve strands. A total of 8 organisms were dissected in order to obtain enough tissue for the lower limit of the RNeasy kit tissue mass specifications. Each dissected tissue was transferred to a microcentrifuge tube, chilled on ice, containing the first buffer of the RNeasy kit, which contains RNase inhibiting guanidine salts. Total dissection time for 8 organisms was around 45 minutes. Tissue was homogenised using a pestle drill piece and variable speed laboratory motor (TRI-R Instruments, Inc.) and total RNA extracted as per manufacturer's guidelines. Total RNA extracted was rapidly utilised for the next steps, to minimise degradation time.

2.2.3 mRNA enrichment of total RNA

mRNA was enriched using the Poly(A)Purist™ Kit [Life Technologies] as per manufacturer's guidelines. 1 mg Total RNA isolated with RNeasy Mini Kit [Qiagen] was used for purification and 0.2 µg of the resulting enriched RNA used in first strand cDNA synthesis.

2.2.4 SMART cDNA synthesis for high throughput pyrosequencing

cDNA synthesis was conducted using the SMART cDNA Synthesis Kit [Clontech Laboratories, Inc.]. For first strand synthesis 1 µg of total RNA or 0.2 µg of mRNA enriched RNA was added to 1 µl 5' SMART II A primer [AAG CAG TGG

TAT CAA CGC AGA GTA CGC GGG], 1 µl 3' RACE CDS [AAG CAG TGG TAT CAA CGC AGA GTA $C(T)_{30}V$ N] and water to 5 µl total volume and heated to 70°C for 2 minutes. 2 µl 5x First Strand Buffer, 1 µl dNTPs, 1 µl dTT (20mM) and 1µl of PrimeScript Reverse Transcriptase [Takara] was added and incubated at 42°C for 90 minutes. Second strand synthesis utilised the Advantage 2 PCR kit [Clontech Laboratories, Inc.], as recommended in the manufacturer's manual, with the SMART Nested Universal Primer (NUP) [AAG CAG TGG TAT CAA CGC AGA GT]. The resulting cDNA was sequenced at Food and Environment Agency (FERA) using a Roche 454 GS FLX Pyrosequencer.

2.2.5 Rapid Amplification of cDNA Ends (RACE)

RACE experiments were performed on 1 µg of total RNA using SMART RACE cDNA amplification kit [Clontech], according to the manufacturer's protocol. RACE experiments were used to retrieve the complete 5' and 3' end of partial cDNAs, including any untranslated regions (UTR). Gene specific primers (GSP) from partial cDNA sequences were designed, a sense primer was used in 3' RACE experiments and an antisense primer in 5' RACE experiments.

2.2.6 cDNA synthesis with random hexamers

cDNA for degenerate PCR primed with ion channel primers was produced using the Transcriptor High Fidelity cDNA Synthesis Kit [Roche] following manufacturer's protocol, following the protocol steps for using random hexamers. 1 µg of total RNA extracted from neuronal tissue using the RNeasy kit method described previously (2.2.2) was used as the RNA input for the cDNA synthesis.

2.3 Molecular cloning

The general cloning protocol was as follows, deviations to this method are identified in specific cloning methods. PCR products were purified from either from Agarose gels using QIAquick Gel Extraction Kit [Qiagen] or Phenol-chloroform ethanol precipitation. Both PCR products and target plasmid vector were restriction digested with relevant restriction enzymes and the best suitable buffer for double digestion. The PCR insert was restricted such that 5' and 3' overhangs

complemented equivalent overhangs of the linearized plasmid vector. Restriction digestions of plasmid vector and PCR insert were checked on 0.8% Agarose gel for correct size and linearization. Where restrictions were done for the first time or with new enzymes, single digestions were done as a control to diagnose any inefficiency in digestion. The digested products were separately purified with phenol-chloroform ethanol purification, the precipitant re-suspended in a small volume of water (10-20 µl depending in the observed quantity in the previous Agarose gel assay) and assessed by gel electrophoresis and quantified with NanoDropTM Spectrophotometry.

The digested products were diluted to concentrations of a 1:1 molar ratio (calculated by equal mass relative to size ratio) and were combined with T4 DNA ligase and buffer [Promega] as following manufacturer's guidelines. Ligation mixture was chloroform extracted and then 1 µl added to 60 µl electrocompetent *E. coli* TOP10 from glycerol stock and transformed as per previously described method. The cell ligation mixture was added to a 500 µl microcentrifuge tube of LSLB and incubated at 37°C for 10 minutes or 1 hour depending on antibiotic, inhibitory or bactericidal respectively. 1-10% of the incubated medium was then plated out per 25 ml Agarose gel petri dishes with a relevant antibiotic to select for the transformed colonies. For plasmid DNAs allowing blue/white screening LB-agar was supplemented with 40 µg/ml 5-bromo-4-chloro-3-indoyl-β-D-galatoside (X-Gal) and 0.1 mM isopropyl-β-D-thiogalatoside (IPTG).

Transformations were conducted at the end of the day and the resulting plates incubated at 37°C overnight and taken out in the morning. A selection of colonies were picked, preferentially larger ones, and were screened by colony PCR using PCR conditions used for the original PCR, alongside the PCR product as a positive control and empty vector as a negative. Several of the resulting positive clones were cultured overnight by adding cells from the transformation plate to 10 ml McCartney bottles of LSLB and incubating 37°C. Plasmids were then extracted as previously described and plasmid DNA was then sequenced by an Applied Biosystems 3730 capillary sequencer in the 5' and 3' direction using relevant primers gene specific primers. The resulting sequences were aligned to the expected sequence used to design the primers

and checked for errors using the Sequencher™ Version 4.5 (Gene Codes Corporation) application.

Cathepsin 5' Fragment	
Cath_L_5_For_EcoRI	GCA**GAATTC**TTCTAGAAGACACCGTCTGGTTATC
Cath_L_5_Rev_XbaI	TGC**TCTAGA**GCCTTGAAATGTTGGCCTTCC
Cathepsin 3' Fragment	
Cath_L_3_For_EcoRI	GCA**GAATTC**ACTGAGGTCAGCTACCCCTAC
Cath_L_3_Rev_XbaI	TG**CTCTAGA**CGCACGATGGGGTAGCTTGCTTGTG
Cathepsin qPCR Primers	
qPCR_5'_CATH_FOR	TTCAAGGCCACTGGAAAACTG
qPCR_3'_CATH_REV	CCGAACTTCTGTGAGCAATCG

Table 2: Cathepsin L RNAi construct primers

Region in primer sequence highlighted refers to restriction enzyme recognition site. All primers are in 5′ → 3′ direction.

2.3.1 Cloning cathepsin L dsRNA construct

First strand cDNA was generated from 1µg total RNA extracted from digestive gland of *D. reticulatum* as previously described and 0.1 µl used as a template for 2 high-fidelity PCR reactions with primers designed with EcoRI and XbaI restrictions in, see Table 2. The resulting products were 415 bp & 419 bp for the 5' and 3' fragments respectively. These were then purified by QIAquick Gel Extraction Kit [Qiagen] and restriction digested with EcoRI and XbaI, an equivalent digestion with the same enzymes was setup for the pLitmus28i vector. The resulting digested products were checked by gel electrophoresis for complete digestion then purified as before. Screening and selection of colonies was done as described previously. This resulted in 2 pLitmus28i-cathepsin L fragment constructs, which were then linearised twice in both directions via restriction digestion. This resulted in

4 *in vitro* transcriptions, a 5' fragment: forward and reverse strands and a 3' fragment: forward and reverse strands. Primers used for qPCR of the gene are included in Table 2.

2.3.2 Cloning apoptosis dsRNA construct

First strand cDNA was generated from 1 µg total RNA extracted from whole tissue of *D. reticulatum* as previously described and 2 µl used as a template for high-fidelity PCR reaction to amplify a region similar to an apoptosis inhibitor from contig C4281;N-;pS1972d. Primers used are shown in Table 3 and the resulting PCR product was of length 342 bp. Only small amounts of PCR product was produced, so the product was cloned into pJet2.1 vector using the CloneJET^{TM} PCR Cloning Kit [Fermentas], rather than restriction digested and cloning into pLitmus28i. pJet2.1 includes a T7 promoter so can be used as the source plasmid for dsRNA, 2 of the clones screened were in reverse orientation, so this construct was used for *in vitro* transcription. This resulted in no necessity to create further primers with restriction sites. Product was cloned into pJet2.1 in both directions and then linearised with XbaI and XhoI which are closest sites to the insertion point. These were then used for dsRNA production in the same manner as the linearised Cathepsin L-pLitmus28i constructs. Construct and qPCR primers are shown in Table 3.

dAPIN RNAi Primers	
5' Apoptosis	AGTCGAACACCGGAACCACTACCC
3' Apoptosis	TGGCGTGCTCCGTCCAAGG
dAPIN qPCR Primers	
5' Apoptosis qPCR	TGGGAGGCTAGCGACTCTGT
3' Apoptosis qPCR	TCGCACTTTCCGTTGATGAG

Table 3: Apoptosis Inhibitor RNAi Construct and qPCR primers

2.3.3 Cloning GAPDH dsRNA constructs

First strand cDNA was generated from 1 μg total RNA isolated from digestive tissue of *D. reticulatum* as previously described and 0.1 μl used as a template for high-fidelity PCR reaction to amplify a region of GAPDH. Primers used are shown in Table 4, and produced a PCR product of 407 bp. The cloning protocol was as described in cloning of cathepsin L, except restriction enzymes used were XbaI & PstI. For the production of the *in-vivo* dsRNA a second construct was cloned using the same method as pLitmus28i but using the L4440 vector, for the experiment

GAPDH RNAi Primers	
5' GADPH_4RNAi_XbaI	GA**CTGCAG**ATATGGATACAGCAACCGGG
3' GADPH_4RNAi_PstI	TG**TCTAGA**ACCTTTCTTAGATGGGTGCC

Table 4: GAPDH RNAi construct primers

Region in primer sequence highlighted refers to restriction enzyme recognition site. All primers are in 5′ → 3′ direction.

described in publication (Solis et al. 2009). Both GAPDH in pLitmus28i and L4440 were cloned into TOP10 cell cultures as previously described. GAPDH-L4440 construct plasmids were then extracted from TOP10 cells using Wizard® *Plus* SV Minipreps DNA Purification Systems [Promega] and plasmid DNA used to transform *E. coli* HT115 (*rnc14::ΔTn10*) cells. This transformation followed the same protocol as described for transformation of *E. coli* TOP10 with the additional use of Tetracycline in addition to carbenicillin antibiotic, in the culture media, as described in publication cited (Solis et al. 2009).

2.3.4 Cloning dTNF protein expression construct

The *D. reticulatum* TNF-L (dTNF) sequence without the transmembrane domain was amplified with high-fidelity PCR from 2 μl of second strand cDNA generated with SMART cDNA synthesis kit [Clontech] from 1μg total RNA isolated

from digestive tissue of *D. reticulatum* extracted as previously described. PCR products could only successfully be amplified from second strand cDNA, possibly due to low levels of transcript available, which lead to a higher PCR error rate. PCR products were then purified by QIAquick Gel Extraction Kit [Qiagen] and, alongside PET32a plasmid vector [Novagen] were restriction digested with NcoI and XhoI restriction enzyme sites, which were included in PCR primers, see Table 5. The resulting digested products were checked by gel electrophoresis for complete digestion then extracted with QIAquick Gel Extraction Kit [Qiagen]. *E. coli* TOP10 cells were transformed with the construct and screened via PCR as described previously. The sequence was compared with C974d contig for errors. The final clone included one incorrect base which was mutagenised via a Site-Directed Mutagenesis protocol using primers, see Table 5. PET32a-dTNF plasmid with corrected base was then extracted as previously described and plasmid DNA used to transform electrocompetent *E. coli* Origami B(DE3) expression strain, transforming with electrocompetent protocol as described previously.

dTNF Cloning Primers	
5' TNF Ligand	GA**CCATGG**AAAACAATGAGTCACAGG
3' TNF Ligand	TT**CTCGAG**CACAAGGGACTCTGCTAC
dTNF Mutagenesis Primers	
5' TNF_MUTA	AAGCTATTTTGGATTAGTGATCC
3' TNF_MUTA	GACTCTCTTTCCAAGTTGACC

Table 5: dTNF cloning and mutagenesis primers

Region in primer sequence highlighted refers to restriction enzyme recognition site. All primers are in 5′ → 3′ direction.

2.3.5 Site directed mutagenesis of dTNF

Site directed mutagenesis was performed by designing two primers which covered the region which included the incorrect base. Primers were flush with each other but did not overlap, with the forward primer being 3' of the reverse primer,

allowing the whole plasmid to be amplified. The forward primer contained the correct base which would mismatch with the incorrect base in plasmid. The PCR was conducted with standard Phusion HF polymerase [Fermentas] and reagents and 20ng of plasmid purified using Wizard® *Plus* SV Mini-preps [Promega]. The reaction conditions were identified, after optimisation, as Step1: 98ºC at 1min; 35 cycles{ 98ºC at 20sec, 59ºC at 30sec, 72ºC at 7min}; Step3: 72ºC at 10mins. A total of 200µl of PCR product was pooled and purified by Phenol-Chloroform extraction and Ethanol Precipitation. The purified DNA was re-suspended in 40 µl of water and treated with T4 Polynucleotide Kinase (NEB) (20U) in Ligase Buffer (5 µl) to a final volume of 50 µl and incubated at 37ºC for 30 minutes. 3µl of reaction mixture was taken out before adding 30U of T4 DNA Ligase and incubated overnight at 4ºC. The ligation mixture and the 3µl saved were then chloroform extracted and transformed into TOP10 cells. The mixture saved before addition of Ligase was used as a control to show whether cells were transformed only with Ligase present.

2.4 dsRNA Production

2.4.1 *in-vitro* dsRNA Production

in-vitro dsRNA was produced with either the Megascript® T7 kit (Life Technologies) or T7 Ribomax® Express RNAi system (Promega). Inserts were cloned into Litmus28i vector and linearised 5' and 3' of the insert for single stranded RNA production. After single stranded (ssRNA) was produced using manufacturer's protocol and purified using phenol-chloroform and ethanol precipitation, sense and antisense strands were annealed by heating to 80°C in a water bath and then allowing to cool to room temperature over 3-4 hours. dsRNA was confirmed via Agarose gel electrophoresis.

2.4.2 *in-vivo* dsRNA production

in-vivo dsRNA was produced using a protocol based on (Solis et al. 2009) using L4440 vector with RNase III deficient *E. coli* HT115 (*rnc14::ΔTn10*) to express dsRNA *in vivo*. *D. reticulatum* GAPDH gene fragment and Kanamycin control gene were inserted into the L4440 vector using XbaI & PstI. Vectors were

transformed first into TOP10 electrocompetent cells. Colonies were screened and checked by sequencing before being grown and plasmid purified, as previously described, and transformed into HT115 (*rnc14::ΔTn10*) electrocompetent cells. Cells were grown, with 10 µg/µl Tetracycline (HT115) and 50 µg/µl Carbenicillin (L4440), overnight at 37°C in 10ml LSLB cultures. 1ml was transferred to 100 ml LSLB with described selection and grown at 37°C until 600 nm OD of ~0.5A (3-5hrs). For IPTG induced cells, IPTG was added to a final concentration of 2 mM and grown at 37°C for 4 hours. Cells were centrifuged at 4000 g for 30 minutes at room temperature and supernatant discarded. Cell pellet was resuspended in 1 M ammonium acetate, 10 mM EDTA at 2% of the original culture volume and phenol-chloroform extracted and Ethanol precipitated as described previously. Nucleic acid pellet was resuspended with 8U Turbo DNase (Ambion) and 40 µg of RNase A to a total volume of 400 µl with nuclease free water. GAPDH dsRNA required further optimisation, with 25% of RNase treatment needed for Kanamycin showing best results, this was equal to 10 µg of RNase A to a total volume of 400 µl. 1 µl of nucleic acid was saved before adding nucleases for comparison. Nuclease treatment was incubated at 37°C for 30 minutes and then phenol-chloroform extracted and precipitated as in previous steps. Nucleic acid was resuspended in relevant buffer for the application.

2.5 Quantitative PCR methods

All quantitation values regard the Ct (Cycle threshold) as defined by the StepOne^{Tm} RT-PCR System (Applied Biosystems) software. Additionally the quantitative real-time PCR referred to throughout will simply be abbreviated to qPCR rather than RT-PCR or RT-qPCR.

2.5.1 Total RNA extraction for qPCR

Individual organisms or pooled groups were flash frozen in liquid nitrogen and ground to powder with a mortar and pestle pre-chilled in liquid nitrogen. Where digestive gland tissue was used, digestive glands were dissected from organisms and ground as with whole tissue. As with RNA extraction for cDNA, digestive gland

dissection took approximately 5 minutes. 50-100 mg powder was transferred to Tri-Reagent (Sigma) and RNA was extracted according to manufacturer's protocol. RNA was resuspended in nuclease free water, and incubated with 1U Turbo DNase at 37°C for 30 minutes. RNA was purified using Phenol-chloroform and Ethanol precipitated and quantified with Nanodrop1000 Spectrophotometer.

2.5.2 Quantitative PCR (qPCR)

1 µg RNA was used in a first strand cDNA synthesis using Nanoscript RT kit [PrimerDesign Ltd.] following manufacturers protocol. The resulting cDNA was quantified as before and unless otherwise stated 40 ng of cDNA was prepared with Precision One-Step qPCR 2x Master Mix with ROX (PrimerDesign Ltd.) and qPCR primers at 0.5 µM final concentration. qPCR primers were designed using Primer Express [Applied Biosystems] and ordered as custom oligonucleotides from Sigma-Aldrich. The qPCR mixture was performed and analysed with a StepOne^{Tm} RT-PCR System [Applied Biosystems].

2.5.3 qPCR Analysis

All qPCR data shown uses comparative CT with relative quantitation values representing $2^{-\Delta\Delta Ct}$, the difference of HKG gene to target gene relative to an arbitrary reference value usually the sham control. Error bars all represent 1 standard deviation of ΔCt as +/- addition to the ΔΔCt exponent.

2.5.4 HKG analysis

Housekeeping gene (HKG) analysis was conducted similarly to previously described study in Mollusca (Sirakov et al. 2009). Primers were designed for contigs which showed strong homology to known HKG and were double checked by hand, see Table 6. In addition to using geNorm (Vandesompele et al. 2002), Bestkeeper (Pfaffl et al. 2004) and Normfinder (Andersen, Jensen, and Ørntoft 2004) the comparative delta-Ct method was also used (Silver et al. 2006). Merged data was produced using the online Reffinder tool (Xie et al. 2012). All qPCR runs including 3 technical replicates for each sample primer combination.

Actin (C2471;N689;S1995d)	
5' Actin_qPCR	GGAAGGATGGCTGGAACAAAG
3' Actin_qPCR	CGGACAGGTCATCACCATTG
EF1-A (C1490;N398;S553d)	
5' EF1-A_qPCR	TGTGCGTGGAGTCCTTCCA
3' EF1-A_qPCR	CAGTCTGCCTCATGTCACGAA
GAPDH (C2454;N37;S199d)	
5' GAPDH_deroc_For	GGCATCGTTGAGGGTTTGAT
3' GAPDH_deroc_Rev	TGCTGGGTCCATCTACAGTCTTC
Tubulin (C2968d)	
5' Beta-Tubulin	GGAACCTTTTAAGCGAGTTGGA
3' Beta-Tubulin	CAGTGTACCAATGCAAGAATGCA
Ubiquitin (C408d)	
5' Ubiquitin_qPCR	CGGCATTAAGAGAGAACTCAAAGTG
3' Ubiquitin_qPCR	GGCAACACCTCCTTAAGCTCAT

Table 6: Primers for housekeeping gene qPCR primers

Table contains a list of housekeeping genes and primers designed for them. The largest tubulin and ubiquitin contigs were only found in CLCBio but have strong BLAST and INTERPRO homology indicating they are correct predicted proteins.

2.6 Protein Expression Methods

2.6.1 Calculating Protein Molecular Mass

Protein mass was calculating using the online Compute pI/Mw tool on the Swiss Institute of Bioinformatics Expasy Server (Gasteiger et al. 2005).

2.6.2 Protein Extraction Method

Protein was extracted from expressing cultures by sonication in 50ml Falcon tubes kept on ice, using a Soni-Prep 150 [MSE] at 26 microns. Optimal sonication

time was calculated by sonication assay comparing time points with protein concentration. Protein concentration was calculated both by CBB gel with ImageJ calculation and bovine serum albumin (BSA) assay. Optimal regime chosen was 16x 30 second sonications interspaced with 2 minute periods to allow cooling. The time is relatively long most likely due to the age of the sonication machine and volumes sonicated; temperature in the cell suspension was measured and never exceeded 15°C.

2.6.3 Protein Purification with HisTrap column

A 5 ml HisTrap HP column [GE Healthcare] was used for purification of Trx-dTNF. Uninduced soluble fractions were equilibrated in binding buffer to a final concentration of 20 mM Na_2HPO_4, 400 mM NaCl (pH 7.4 with HCl) and loaded onto His-Trap column, which was pre-equilibrated with binding buffer (BB), at 5ml/min at room temperature. Later loading, which was used for injection assays and gels, was done with BB containing 10mM Imidazole. Column was run for 2x protein fraction final volume and then washed with BB until OD came to a baseline on a FPLC chart recorder. This was repeated with BB containing 25 mM initially and 50 mM in later purifications of Imidazole, the elutions of which were saved for comparison. The protein was eluted with BB containing 200 mM Imidazole until OD had peeked and had almost returned to baseline level on the FPLC chart recorder.

2.6.4 Dialysis & lypholization

Fractions eluted from HisTrap columns were pooled and after confirmation on CBB gels and initially western blot, fractions were dialysed against 20 L distilled water with ~2 g of 50 mM ammonium hydrogen bicarbonate as an anti-microbial agent. 12-14 KDa dialysis tubing was prepared by boiling in a solution of 2% (w/v) sodium bicarbonate and 1 mM EDTA (pH 8.0) for about 20 minutes. Dialysis was carried out at 4°C with constant stirring and 6 changes of water every 1 hour except 1 change which was left overnight. Dialysed protein samples were frozen on the walls of freeze-drying flasks by shelling in liquid nitrogen and lyophilized on a vacuum freeze-dryer overnight.

2.6.5 Buffer Exchange Column Purification

Vivaspin 20 [Sartorius] columns were used to purify protein from HisTrap column eluent. Fractions from HisTrap column were treated according to manufacturer's guidelines. 10 ml of HisTrap elution was buffer exchanged with TAE buffer (pH 7.4) and centrifuged at 4000 g over the course of 2 days, with column kept overnight at 4°C. Buffer was refreshed 8 times in total, each refresh after the total volume of sample had reach 200 µl. Centrifugation step was done over a long period of time due to formation of precipitate which impeded the flow through of buffer. The last 3 buffer flow-through pH matched the buffer pH of 7.4, indicating the buffer had been fully exchange. The resulting 200 µl of sample in TAE was transferred to a new tube, avoiding transfer of precipitate, and protein confirmed with CBB and western protein gels.

2.6.6 BCA Assay

Total protein was quantified using BCA protein quantification kit (Pierce) and BSA (Bovine serum albumin; 0.1-4 µg/ml) as a standard. Concentrations of unknown protein were predicted using the standard curve. BCA reagent was prepared by mixing Solution B with A (1:50) freshly before use. In microtitre plates 10 µl of each standard or unknown sample was added to separate wells (in triplicate) and then mixed with 200 µl of BCA Reagent. The plate was incubated for 30 minutes at 37°C. Absorbance was then read measured at 562 nm using VERA max microplate reader [Molecular Devices].

2.7 *Deroceras reticulatum* Methods

2.7.1 *Deroceras reticulatum* sourcing

D. reticulatum were collected from the wild, on and around the grounds of the Department of Food, Environment and Rural Affairs (FERA), Sandhutton, York, YO41 1LZ. These were either kept in culture conditions described in the following

sections at FERA or transferred and kept in similar conditions at the University of Durham.

Figure 5 : **Deroceras reticulatum**

Deroceras reticulatum *individuals feeding on gem lettuce. Area ringed in red is a 'lesion' on the surface of an individual, the cause of which is not clear. Would have a wound-like appearance with a light discoloured area on the surfaced with darker ridge ringing the area. This would indicate the organism would likely perish within the following weeks, as would all others within the culture container.*

2.7.2 Maintaining cultures

Cultures were maintained until organisms were utilised in various experiments described. *D. reticulatum* individuals were kept in plastic containers containing moistened paper toweling on the base and a selection of lettuces as well as wheat germ as a food source. Organic lettuce was either grown by FERA or sourced from local suppliers, organic wheat germ was sourced from local suppliers. Cultures were moved to new containers weekly and given fresh food. Cultures were incubated at 10°C, high humidity, with a 16 hour day/night cycle.

D. reticulatum cultures were prone to disease and were observed to form 'lesions' which would quickly spread to other organisms within a container, (See

Table 5). In these cases organisms were removed from the containers and not utilised further for any experiments. Despite this several cultures were maintained through multiple generations with eggs collected, washed in a very mild bleach solution and incubated in petri-dishes containing moistened paper toweling and incubated in the same conditions as adults until hatching when they were moved to plastic containers. Eggs took approximately 1 month to hatch, but varied largely with some eggs taking up to 3 months to hatch. In some cases eggs would turn red over time for an unknown reason and would fail to hatch.

2.7.3 Injection Assays

Injection assays were conducted by first anaesthetising *D. reticulatum* on ice and then injecting through a central point of the foot with the needle pointing toward the posterior. Injections were done with a 5µl, 10µl or 20µl Hamilton syringe which were washed well with both ethanol and relevant buffer between injections of individual organisms. Dye injections were done with 14µl of PBS and 1µl Dr. Oetker SuperCook Blue Food Colouring Dye containing Brilliant Blue FCF.

2.8 Bioinformatic Methods

2.8.1 Sequencing

Classic Sanger sequencing was carried out using BigDye Terminator with AmpliTaq DNA polymerase (ABI Biosciences). Reaction products were analysed on automated sequencer, ABI Prism 377 XL, DBS Genomics, Dept. at Durham University, School of Biological and Biomedical Sciences. Plasmids inserts were completely sequenced on both strands of the DNA by using primers directed against determined sequence to complete overlaps.

454 Sequencing of cDNA was conducted using the automated throughput of a 454 GS FLX platform, by The Food & Environment Agency (FERA) (Sand Hutton, York, YO41 1LZ).

2.8.2 Univec Cleanup

Sequences were input into BLASTn using settings as used by VecScreen online tool, to guarantee the same output a BLAST strategy file was generated from VecScreen and used with BLASTn. Matches were assessed using VecScreen criteria with terminal matches define as with 25bp of the beginning or end of the sequence. Any match type (weak, moderate, strong, suspect) was removed using custom Java classes.

2.8.3 Sequence Data Assembly

After Univec cleaning any reads <50bp after trimming were removed. The reads were then assembled with the following assemblers: Newbler v2.6, SeqMan NG v4.1.2 and CLCbio 4.7.2.

Newbler v2.6 was run using the GS De Novo Assembler graphical interface with the following parameters: seed step = 5; seed length = 16; seed count = 1; minimum overlap length = 40; minimum overlap identity = 90; Alignment identity score = 2; Alignment Difference Score = -3; rip = 1.

CLCbio 4.7.2 used standard parameters: -conflict vote; -non_specific random; -paras Default. SeqMan NG used the following parameters: Match Size = 21; Match Spacing =75; Minimum Match Percentage = 85; Match Score = 10; Mismatch Penalty = 20; Gap Penalty = 30; Max Gap = 15; Genome Length = 1674712; Expected Coverage =21.

Reads and contigs were then uploaded to a BioSQL based MySQL database to allow easy comparison. Statistics for assemblies were generated using the respective ACE files. The terms n50 and n90 may have differing definitions, the algorithm used was based on the source code of the Mauve multiple genome alignment software.

2.8.4 Dataset Upload

14118 CLCbio contigs and unassembled reads were uploaded to Genbank Transcriptome Shotgun Assembly (TSA) database, TSA accession numbers JW036070 – JW050187. 5332 sequences (563 of which were contigs) were not

uploaded due to size being below the 200bp threshold for TSA submissions. 6 further unassembled reads were removed due to having >10% Ns.

2.8.5 BLAST Homology

Contigs and unassembled reads were analysed by comparison with the NCBI non-redundant protein database. No minimum e-value cutoff was used as cutoffs were applied post-run through database filtering. As majority of the processing time is during sequence comparison, there's no significant improvement in performance from using higher cutoffs. Hit quantities were limited to 50 hits, to reduce the overhead of manipulating larger datasets. The homology searching was conducted using BLASTx from the NCBI BLAST 2.2.25+ standalone package. BLAST jobs were run against the none-redundant (nr) protein database across a number of standard personal computers and servers. BLAST XML data was uploaded to the BioSQL database using a combination of BioJava and woodstox class library (http://woodstox.codehaus.org/) for XML for parsing BLAST data.

2.8.6 Peptide Prediction & InterProScan

Predicted polypeptides for assembled sequences were produced using ESTScan2.1 (Iseli et al. 1999). The resulting amino acid sequences were then analysed by InterProScan (Zdobnov & Apweiler 2001) which uses 14 separate software packages to analyse the peptide sequence for peptide motifs and homologous domains.

2.8.7 Phylogenetic Analysis

NCBI hits were mapped to phylum by retrieving taxonomy data for Gene Identities (GI) numbers using the NCBI Entrez query service. Phylogenetic trees for individual genes were generated using Clustalw2 with neighbour-Joining method and a bootstrap value of 1000; and visualised using the Forester Java class libraries (PhyloXML (Han and Zmasek 2009)). Where genomes were available using NCBI Map Viewer (http://www.ncbi.nlm.nih.gov/projects/mapview/) with BLASTx against each species geneset or refseq databases with *D. reticulatum* predicted peptides. Where Mollusca and other species without Genomes were included, these sequences

were either retrieved from NCBI nr via BLASTx, or derived from EST data using tBLASTx against EST_OTHERS NCBI database.

2.9 Biology Database

For many of the bioinformatic methods, processes were automated and many of the statistical analyses were conducted by uploading data into a database and filtering or mapping data. Tools used to manage the scripting and database uploading are included in a biology toolkit, written in Java programming language and available at https://github.com/EnderDom/Eddie. Database schema used was BioSQL and can be found at http://www.biosql.org. Data was input through either the custom scripts including in the biology toolkit or with scripts from the bioPerl library. BioSQL supports a number of different database types, MySQL database was used for this project with a number of scripts being specific to that database type. Tables were added on top of the base schema to accommodate information about multiple assemblies as well as inclusion of additional 'run' information which kept track of the software versions, parameters and dates of the various data generated. However the base schema was not altered in anyway such that it should be in theory compatible with any other software the implements a BioSQL interface. Majority of the statistics generated from the dataset were done without any scripting/programming required, purely with SQL database queries.

2.9.1 Data upload

Read sequences were uploaded to BioSQL from assembly files of the ACE format produced by the assembly software. The ACE files were parsed with a custom ACE file parser written in Java, a custom name identifier in the format Assembler_Tissue_Number was used along with ids given by the assemblers in the database bioentry ID. The database bioentry ID is default indexed and used to link to other tables, including the biosequence table which holds sequence information. The contig data was uploaded in the same way, and reads mapped to contigs by adding contig and read ids as rows to an assembly table. The assembly table was not part of the base BioSQL schema, which was not originally designed to hold assembly data.

An additional table name run was also added, this was another custom table which holds information about the software used to produce the assemblies (and other data uploaded such as BLAST and INTERPRO) as well as dates run, version and parameters used. BLAST data was uploaded via parsing BLAST XML files produced as output of the BLAST homology searches, using the Woodstox Xml parser. Key data such as accessions, start/end, score and evalue was then transferred to the dbxref database tables. All data was added as columns to the dbxref_bioentry linker table rather than attempting to alter the seqfeature/location tables. INTERPRO data was uploaded in a similar manner to BLAST data but used the term tables to store IPR terms with specific software matches such as pfam and panther accessions being added to the dbxref tables and linked through the term_dbxref table to linked IPR terms. IPR terms were linked to global IPR terms such as binding site, active site, domain etc. through the term_relationship table which enabled selection of subsets such as all active site IPR terms.

2.9.2 Taxonomy Mapping

Accessions were mapped to NCBI taxonomy IDs using the e-utilities HTTP API (Federhen 2011) which enables programmatic access to the NCBI database through URL base queries. NCBI accessions were retrieved from dbxref table in the BioSQL database and used to retrieve the taxonomy ID for each accession from the NCBI database. Where the taxonomy has not already been retrieved, it was downloaded using e-fetch and uploaded into the taxon and taxon_name tables of the BioSQL database. Once all taxon data was downloaded left and right values were calculated for the nested-set representation of the taxonomy data. This allows for selection of entire child groups from the parent node taxon ID using very simple SQL queries (Mackey 2002).

2.9.3 Assembly & Homology Metrics

The majority of data statistics demonstrated were calculated using SQL queries against the modified BioSQL database. These were executed either with a standard MySQL command line client or using the phpMyAdmin software package

hosted on an apache2 server. BLAST evalue graph data was produced using SQL query:

```
SET @runtot:=0; SELECT q1.EVALUE, (@runtot := @runtot +
COUNT) AS CUMULATIVE FROM (SELECT bioentry_dbxref.evalue AS
EVALUE, COUNT(bioentry_dbxref.evalue) AS COUNT FROM
bioentry_dbxref WHERE  bioentry_dbxref.run_id=x AND hit_no=1
AND bioentry_dbxref.rank=1 GROUP BY bioentry_dbxref.evalue
ORDER BY bioentry_dbxref.evalue;) AS q1
```

bioentry_dbxref.run_id is set to the relevant BLAST run. This outputs a cumulative count for each BLAST hit with its evalue, which can be used as x, y coordinates. Species data was produced using SQL query:

```
SELECT dbxref.ncbi_taxon_id AS taxid, taxon_name.name AS
taxname, COUNT(dbxref.ncbi_taxon_id) AS count FROM dbxref
INNER JOIN taxon USING (ncbi_taxon_id) INNER JOIN taxon_name
USING (taxon_id) INNER JOIN bioentry_dbxref USING (dbxref_id)
INNER JOIN bioentry_run USING (bioentry_id) WHERE
bioentry_dbxref.hit_no=1 AND bioentry_run.run_id=x AND
bioentry_dbxref.evalue<1e-3 AND
taxon_name.name_class='ScientificName' GROUP BY taxid ORDER BY
count
```

Phylum data required an additional programmatic step, all phyla IDs were selected by using SQL query:

```
SELECT ncbi_taxon_id, taxon_name.name FROM taxon INNER
JOIN taxon_name USING (taxon_id) WHERE taxon.node_rank LIKE
'phylum'
```

These were added to a list and iterated over. For each phylum all species were selected using the nested set representation and the number of top BLAST hits with this species was counted and the data appended to the phyla, using the following query:

```
SELECT COUNT(bioentry_id) AS COUNT FROM bioentry_dbxref
INNER JOIN dbxref USING (dbxref_id) WHERE dbxref.ncbi_taxon_id
IN (SELECT taxon.ncbi_taxon_id FROM taxon INNER JOIN taxon AS
include ON (taxon.left_value BETWEEN include.left_value AND
```

```
include.right_value) WHERE include.ncbi_taxon_id=?) AND
bioentry_dbxref.run_id=x AND bioentry_dbxref.evalue<1e-3 AND
bioentry_dbxref.hit_no<=1
```

Kernel density plots were produced by simple selection of length values of biosequences using run ID to differentiate reads from contigs and different tissue types, such that the query was:

```
SELECT biosequence.length FROM biosequence INNER JOIN
bioentry_run USING (bioentry_id) WHERE run_id=x
```

This query output a basic list of each sequence length from the run identified with x. The kernel density was then calculated and graphed using the inbuilt kernel density function in the R programming language.

The Venn diagram data showing overlap of reads between different assemblies was calculated using the MySQL IN query function. This allows for subsets of data to be selected from a previous query, without the need from any programmatic assistance. For each assembler a query selecting all the read IDs in the assembly, using the assembly table, is constructed.

```
SELECT read_bioentry_id FROM assembly WHERE run_id=1;
```

In this case run_id 1 is the CLCbio assembly for the digestive gland dataset; this selects all the reads from that assembly. To identify how many of those also exist in run_id 2, the Newbler digestive gland assembly, we perform both queries then select all read_bioentry_ids that exists.

```
SELECT COUNT(DISTINCT(read_bioentry_id)) FROM assembly
WHERE read_bioentry_id IN (SELECT read_bioentry_id FROM
assembly WHERE run_id=1) AND read_bioentry_id IN (SELECT
read_bioentry_id FROM assembly WHERE run_id=2)
```

The use of DISTINCT counts only unique reads as the same read is present in both selection statements, as the selection of IDs is not exclusive. It should be noted that this could also be done with a self-join of the table, which may be a slightly better optimised method, though functionally less explicit, i.e.:

```
SELECT COUNT(assembly.read_bioentry_id) FROM assembly
INNER JOIN assembly AS self ON
```

```
(assembly.read_bioentry_id=self.read_bioentry_id) WHERE
assembly.run_id=1 AND self.run_id=2
```

However in both cases the resulting value is the same, the difference in time is less than a second for the current datasets. Each value for each pair and then all 3 assemblers were calculated and Venn diagram drawn with the 'venneuler' CRAN library (Wilkinson 2012).

Chapter 3 | Initial Investigation of Transcripts from *D. reticulatum*

The very first step in the investigation of *D. reticulatum* transcripts is the production of RNA of sufficient quality that complementary DNA (cDNA) products of reverse-transcribed RNA can be produced of sufficient length to be identified as partial or whole transcripts. Optimizing the RNA extraction and cDNA synthesis was the initial stage of this project focusing on producing degradation-free RNA and then good quality cDNA. Figure 7 demonstrates target appearance of both RNA and cDNA with gel electrophoresis.

After RNA and cDNA quality met acceptable standards, random cloning of cDNA was conducted. This was done as a proof-of-principle, that the cDNA seen on gel represented mRNA transcripts found in *D. reticulatum*. Whilst no sequence data was available to compare this data to, homology analysis to transcripts expected to be found in *D. reticulatum* based on previous biochemical work was considered acceptable evidence that the cDNA could produce transcript sequence data. This body of work would serve as evidence that the cDNA and homology analysis of it would produce relevant and useful data were it to be sequenced using high-throughput sequencing technology.

When cDNA has been proven to produce transcript sequences that can be identified through homology analysis, the cDNA can be used to investigate individual gene transcripts. Both considering sequences found through random cloning and targeted approaches using degenerate PCR based on homologous sequences already known. In both cases partial sequences can be extended through rapid amplification of cDNA ends (RACE). RACE produces complete sequences from partial ones by using one gene specific primer (GSP) and one cDNA adaptor primer to polymerise unknown regions of the sequence 5’ and 3’ of the known region. In combination with degenerate PCR this allows complete transcript sequences to be generated from cDNA when a homologous sequence is sufficiently similar for degenerate primers to work.

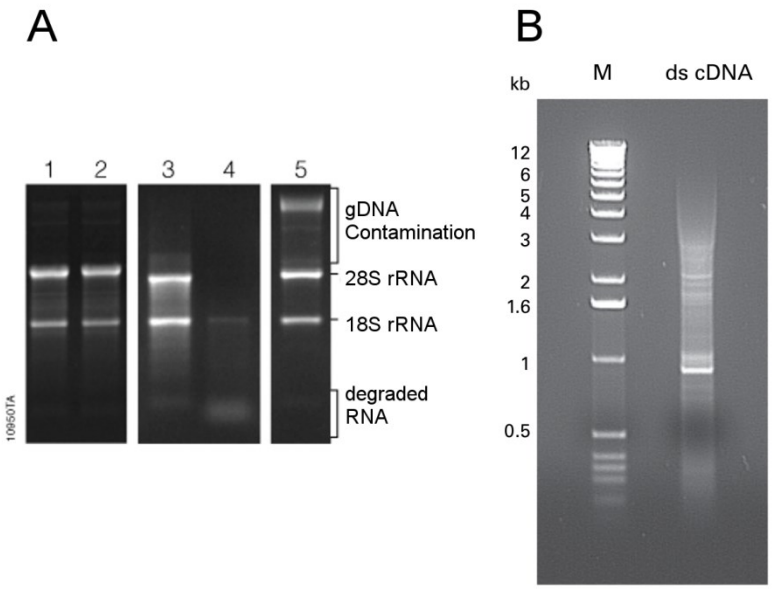


Figure 6 : Example Total RNA and cDNA

Image A is an RNA analysis with agarose gel electrophoresis shown in methods of RNA quality assessment [Promega] (http://www.promega.co.uk/resources/pubhub/methods-of-rna-quality-assessment) [Accessed 04/12/13]. Lanes 1 & 3 show intact RNA with no degradation. Lane 3 shows smearing due to degradation with greater degradation leading to loss of 28S rRNA band in lane 4. Lane 5 includes significant gDNA contamination. Image B represents cDNA shown in the SMART PCR cDNA synthesis kit User Manual [Clontech]. The cDNA here is second strand cDNA after 15 thermal cycles using human placental total RNA as the source RNA. These gel electrophoresis images of RNA and cDNA were used as a comparison to assess visually the quality of cDNA and RNA by Agarose gel electrophoresis.

A high priority target for degenerate PCR and RACE was Molluscan ion channels. Many insecticidal compounds, both synthetic and naturally-occurring, target ion channels associated with neuronal tissue, causing paralysis. Neurotoxic peptides which block ion channel proteins, isolated from the arthropod venoms, have been suggested for use as biopesticides, and recombinant proteins based on spider venom proteins are being developed for their insecticidal activity (King 2007;

Windley et al. 2012). A similar strategy could be applied to *D. reticulatum* ion channels making them an obvious target for the development of new molluscicides.

Conotoxins are neuropeptides extracted from the venom of *Conus* genus of snail, where the primary prey of many species is other Mollusca. Some venom peptides extracted from *Conus* spp., which have structures related to those found in spider venom peptides, have been shown to have molluscicidal effects when injected (Fainzilber et al. 1991). ω-conotoxins and κ-conotoxins block calcium and potassium channels respectively, whilst the δ- and μO-conotoxins groups affect sodium voltage-gated channels (Heinemann and Leipold 2007). The TxVIA conotoxin from *Conus textile* was shown to also have insecticidal activity increasing the appeal of conotoxin biopesticides (Bruce et al. 2011).

Whilst a reasonable amount of sequence information is available for the model mollusc *A. californica*, very little is known with regard to characterisation of genes in crop pest molluscs. Based on the limited sequencing done, the difference between sequences of homologous genes in *D. reticulatum* and *A. californica* appears quite large. Characterisation of ion channels through sequencing of encoding transcripts in a wider range of mollusc species would be of benefit for future work to evaluate mollusc-specific neurotoxins. The long term benefits of isolated Molluscan ion channel sequences from target organisms is the potential to carry out *in vitro* assays, based on expression of recombinant proteins, to test effects of potential channel agonists or antagonists. With maintaining organisms being labour-intensive and difficult in some cases, the availability of an assay to assess feasibility of a protein before large scale work begins is obviously advantageous. In seeking to develop molluscicides targeting ion channels which are effective against *D. reticulatum*, the first stage is sequencing the ion channel genes in this target organism.

3.1 Extraction of RNA and cDNA synthesis

3.1.1 Extraction of RNA from *D. reticulatum*

Different methods for RNA extraction were tried with tissues from *D. reticulatum*, in order to identify a method that would give RNA of good quality for

cDNA synthesis. The large amounts of carbohydrate-rich mucus produced by this organism proved a major problem in producing RNA for subsequent use. Extracted RNA was analysed by Agarose gel electrophoresis, followed by staining for total nucleic acid (see Figure 8). RNA quality was initially assessed based on the presence

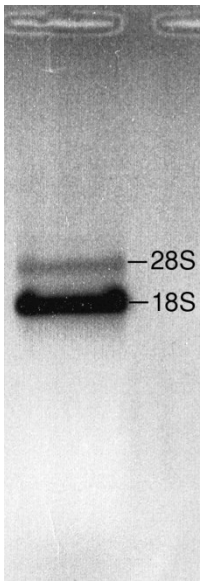


Figure 7 : Gel Electrophoresis of Total RNA with the two main rRNA subunits

Gel electrophoresis of total RNA extracted with Tri-Reagent from whole organism tissue of D. reticulatum. The two main bands visible are the 28S and 18S rRNA subunits of the eukaryotic ribosome. Non-rRNAs are not visible due to relative quantity being significantly lower and size distribution across the gel greater. The expected signs of RNA degradation having occurred is the loss of the larger 28s band and a low molecular weight smear forming below the 18s band.

of both ribosomal bands visible, with no visible smearing which would indicate degradation. The different methods of RNA extraction were then compared by synthesising second-strand cDNA with the SMART cDNA synthesis kit and comparing and assessing cDNA qualities by gel electrophoresis, see Figure 8. The best quality cDNA was that with a clear banding pattern rather than smearing, where the distribution of bands was more evenly weighted across all molecular weights.

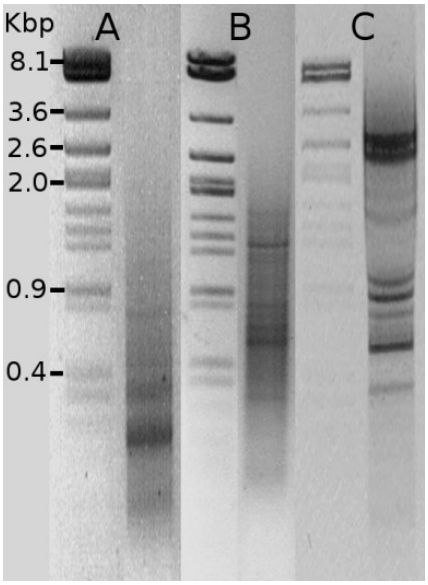


Figure 8 : Gel electrophoresis of D. reticulatum cDNA samples

A,B,C represents 3 separate gels each containing 2 lanes, the first a DNA marker, the second, second-strand cDNA. They have been aligned, based on the molecular weight marker, as a single image for equivalent size comparison of cDNA. A) Whole tissue cDNA extracted with Tri-Reagent, B) Neural tissue cDNA extracted with RNeasy Mini-Kit, C) Digestive gland tissue extracted with RNeasy Mini-kit. A) was an earlier attempt to synthesis cDNA with tri-reagent, and in general can be considered to poorest quality based on size distribution and smearing of the lane. Neural tissue is improved with slightly less smearing and bands at a higher position. Digestive gland is considered the best quality with large bands at around 2.6Kbp. The cDNA here was over-cycled and lanes were overloaded to exaggerate the banding pattern for the cDNAs to better compare them.

Smeared, low molecular weight cDNA was considered inferior and RNA extraction methods producing this cDNA were discounted. Banding pattern of cDNA is expected due to the relative differences in transcription of mRNA. Large bands represent cDNA sizes and likely specific transcripts which are more frequent than others. Fresh tissue homogenisation with electric pestle produced better cDNA than grinding frozen tissue in liquid nitrogen; likely due the improved speed, with hand

grinding being time-intensive due to the size and liquid content of the organism. Overall, selection of specific tissues and the time taken to process the RNA into first strand cDNA appeared to be the major factors in final cDNA quality based on gel electrophoresis assessment. The RNA method with best resulting cDNA was found to be fresh tissue homogenisation with RNA isolation done with RNeasy Mini Kit [Qiagen], but this method was only viable for dissected organs. For RNA extraction from whole organisms, or tissues containing large amounts of mucus, the use of Tri-Reagent [Sigma] was the only viable method; the viscous nature of extracts from whole organisms prepared in the absence of strong denaturants like the Tri-Reagent resulted in the mixture clogging up the RNeasy column based extraction. On the basis of the relatively poor quality of RNA prepared using the Tri-Reagent method, the use of dissected organs as a source of RNA for cDNA synthesis was preferred.

3.1.2 Random cloning into pJET2.1 vector

The digestive gland was considered one of the most informative tissues for investigation, with any potential molluscicides having to pass through the gut. In addition, it can be regarded as a target for molluscicides in its own right, similar to the antimetabolic effects resulting from inhibition of protein digestion by protease inhibitors in insects (Schlüter et al. 2010; P. Pyati et al. 2011). Total RNA extracted with RNeasy Mini Kit from dissected digestive glands of *D. reticulatum* was used as a template for cDNA synthesis, with a SMART cDNA synthesis kit. The resulting cDNA was size fractionated for sequences >1kb and cloned into the plasmid vector pJET 1.2 using a CloneJET PCR cloning kit [Thermo Scientific]. Smaller sequences were removed to try and increase the provision of real transcript sequences as well as to try and counterbalance the blunt-ended ligation cloning bias as smaller sequences are more likely to be cloned into vectors than larger ones. Cloning without size fractionation produced a majority of clones with <100bp insert, which can be too small for effective homology analysis, particularly if fragment represents a less conserved region.

Recombinant vectors were transformed into TOP10 cells and plated on LB agar with carbenicillin (50µg/ml). Plasmids from these clones isolated with Wizard®

Plus SV Minipreps DNA Purification System and assessed for the presence of an insert by colony PCR using pJET 1.2 forward and reverse primers, using the vector alone as a control.

Clone	Size	Summarised most similar Proteins based on BLASTx search	Expect
N17-M7	3kb	-	-
N17-M8	1.7kb	carboxypeptidase, zinc carboxypeptidase A	4e-10
N17-M9	1.8kb	c-type Lectin	4e-10
N17-M15	1.5kb	3-hydroxy-3-methylglutaryl-CoA reductase	1e-10
N17-M18	1.9kb	Snail soma ferritin	4e-34
N17-M26	1.7kb	Snail soma ferritin	1e-33
N17-M28	1.7kb	splicing factor, putative	0.048
N17-M31	1.3kb	-	
N17-M37	1.2kb	-	
N17-M38	2.3kb	tubulin subunit beta	8e-99
N17-M42	2.4kb	cellulase	5e-84
N17-M43	1.6kb	-	-

Table 7: pJET digestive gland sequenced clones and BLAST database matches

Clones were sequenced and vector sequences were removed and insert compared to the BLAST NCBI nr database using BLASTx. The resulting matches were summarised here, matches left blank represent no significant match to a protein from the NCBI nr database. As cloning increases bias for smaller pieces of cDNA, clones were chosen based on their size on gel. This means that they may not be indicative of the prevalent sequences in the transcriptome, but should better represent what genes are present.

A selection of clones were screened for inserts using colony PCR. 50 clones were screened of which 12 plasmids were then selected randomly from those which had observable size difference to empty vectors (>300bp insert) and sequenced in both directions (forward and reverse) by using appropriate primers with an Applied Biosystems 3730 capillary sequencer. The forward and reverse insert sequences were aligned and the complete insert was compared to the global sequence databases using BLASTn and BLASTx homology search software programs. Of 12 clones sequenced and compared, 7 showed sequence similarity to previously characterised sequences in the database, allowing the function of their encoded proteins to be identified, see Table 7.

Sequence similarity to genes whose products were expected to be found in a mollusc digestive gland transcriptome, with the inclusion of several enzymes including a cellulase homologue, showed strong support for the quality of the cDNA and its potential to produce significant gene sequences. The sequences described were subsequently found in the digestive gland pyrosequencing and are discussed further in the next chapter. The data here was used to check the presence of digestive gland sequences and validate the cDNA as source of informative gene sequences.

3.2 Characterisation of full-length transcripts; 5' & 3' RACE with Ferritin Homologue

RACE is technique which uses the same mechanism as cDNA synthesis. 2 sets of first strand cDNAs are produced amplifying 5' and 3' ends of transcripts respectively. One set uses a 5' primer from the known region of the gene ("gene-specific primer") and the common 3' cDNA primer used to prime the reverse-transcription to amplify the 3' end of the gene. The second cDNA uses a 3' primer from the known gene region in conjunction with a common terminal primer complementary to the 5' primer used for cDNA synthesis via template-switching. With this method the full gene sequence can be amplified in two overlapping fragments, allowing a full sequence of the entire transcript to be assembled.

For the RACE reaction ferritin was chosen, due to previous examples of ferritin in Mollusca and the potential link with iron phosphate, a molluscicide

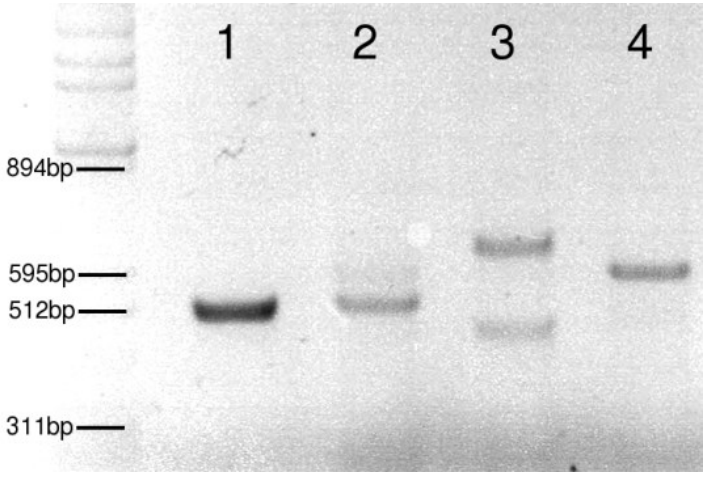


Figure 9 : Gel electrophoresis of RACE PCR products for Ferritin Gene

Gel electrophoresis of PCR and RACE PCR products. Lanes 1 and 2 are PCR products from control PCR between 5′ and 3′ Ferritin gene specific primers (GSP) in the 5′ RACE cDNA and 3′ RACE cDNA respectively. Lane 3 shows PCR product from PCR between 5′ Ferritin GSP and 3′ RACE primer with the 3′ RACE cDNA as template. Lane 4 shows PCR product between 5′ RACE primer and 3′ Ferritin GSP with 5′ RACE cDNA as template. Lane 3 has 2 PCR products, with the smaller product being of lower molecular weight than control product this was assumed to be the result of non-specific binding and amplification. The larger product from lane 3 was isolated via gel extraction and after sequencing deduced to be the correct RACE product.

discussed in chapter 1. Ferritin appeared to be the most interesting gene available to further develop as a potential target for future research. Clones containing a partial ferritin sequence were used as a template for a RACE experiment to attempt to fully sequence a *D. reticulatum* gene RACE cDNA was prepared as with SMART cDNA and 5' and 3' regions were amplified using PCR. 5' and 3' fragments were checked via gel electrophoresis and were of expected length based on homologous sequences, see Figure 10. Amplified fragments were isolated through gel extraction, cloned (into

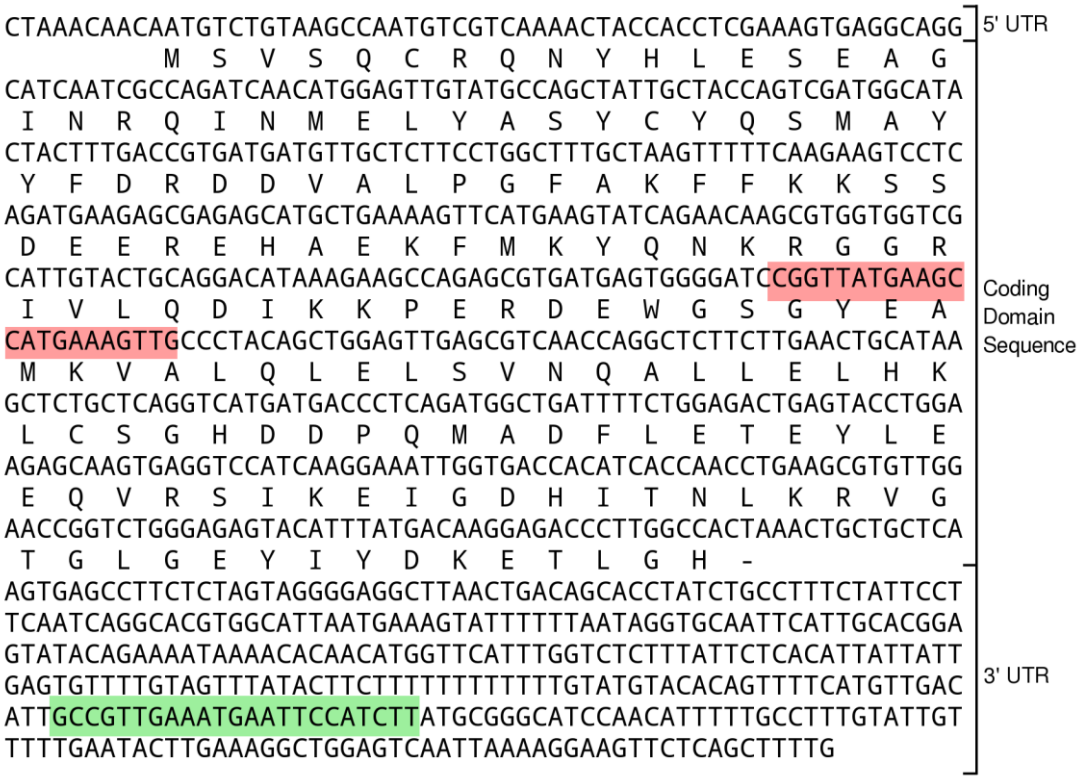


Figure 10 : D. reticulatum sequence homologous to Ferritin identified through 5' & 3' RACE

cDNA sequence for gene homologous to Ferritin. The incomplete sequence was identified via random cloning and sequencing of D. reticulatum cDNA. Primers were designed based on the incomplete cloned sequence, highlighted in the figure 5' (red) and 3' (green) primer sequences. It is likely based on UTR regions of other species that there is still a region of 5' UTR to identify, but the sequence here includes the full coding domain sequence based upon comparison with homologues. This was confirmed later in the project through analysis of ferritin gene in high-throughput sequencing datasets.

pJET 1.2 vector) and sequenced as previously described. Forward and reverse sequences were aligned and compared with known ferritin genes. The resulting 892bp transcript is shown in Figure 11. It contains an open reading frame predicting a peptide of 172 amino acid (aa) residues in length, with a predicted mol. wt. of

20kDa. This transcript was also identified in the high-throughput sequencing, and is discussed in more depth in chapter 4.

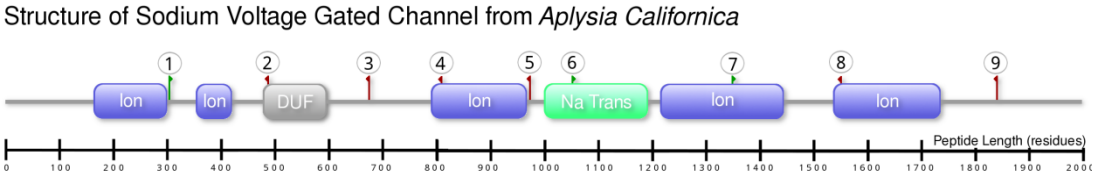


Figure 11 : Structure of A. californica Ion channel with primer locations flagged.

Regions demarcated based on BLAST conserved domains identified in sodium channel alpha-subunit SCAP1 of A. californica (GI: 1842249). Peptide length given in residues, primers marked with flags and numbers corresponding to numbers in Table 8, green flags are 5', red 3'.

DUF: Domain of unknown function (DUF3451); pfam11933, Ion: Ion transport protein; pfam00520, Na Trans: Sodium ion transport-associated; pfam06512

3.3 Ion Channels as a Target for Molluscicides; Degenerate Primer PCR

3.3.1 Overview of Channel homology and PCR design

Figure 12 shows an overview of the domain structure of the sodium voltage-gated ion channel *A. californica.* At the time of this work *A. californica* was the closest species to *D. reticulatum* for which a complete sodium channel gene sequence was available. Highlighted are domains specified by NCBI CDD data for GI: 1842249 which correspond to areas with known function for the sodium channel. Sodium voltage-gated channels ion transport domains correspond to the transmembrane areas which loop through the cell membrane. A multiple sequence alignment of the *A. californica* sodium channel with homologues from other

No.	Name	Sequence	Redundancy	Position (bp)
1	5' Na PWN	CMTGGAAYTGGCTNGAYTT	32	909
2	3' Na QDF	TCCCARWARTCYTGNGTCAT	64	1462
3	5' Na DAWC	TTYACNRAYGCNTGGTGYTGG	256	4034
4	3' Na YIAVI	ATRACNGCRATRTACATRTT	64	5509
5	5' Na WIES	GARTGGATHGANTCNATGTGG	96	3146
6	3' Na VII	TTNTCRATRATNACDCCRATRAA	768	4645
7	3' Na GAIA	ACCATNACRTCNYTCATNTC	256	2424
8	3' Na KLAKS	GGCCAIGAYTTBGCIARYTTRAA	768	2913
9	3' Na AAV	GGNGTNACNGCCCGTGARTCRTC	256	2023

Table 8: Redundant Primers designed based on A. Californica Na voltage gated channel

invertebrate species was made, and used to predict regions of sequence showing the least variability. The alignment is not shown here due to size, but areas of high conservation between species were selected and primers were designed based on all possible permutations of nucleotide sequence based on amino acid sequence. The degenerate primers for PCR are shown in Table 8.

3.3.1 Partial sodium channel sequence

Primers 1-6 were initially assessed with varying source tissue, PCR protocols and cDNA methods. The combination of low annealing temperatures and degenerate primers led to a large amount of non-specific binding. Predicted sizes based on homologous genes were used to screen PCR band products. The only amplification reaction which produced a region of the sodium channel, used neuronal cDNA produced with random hexamers as a template, in combination with a touch down PCR protocol, with primers 1 and 2. Figure 13 shows a gel electrophoresis of touch

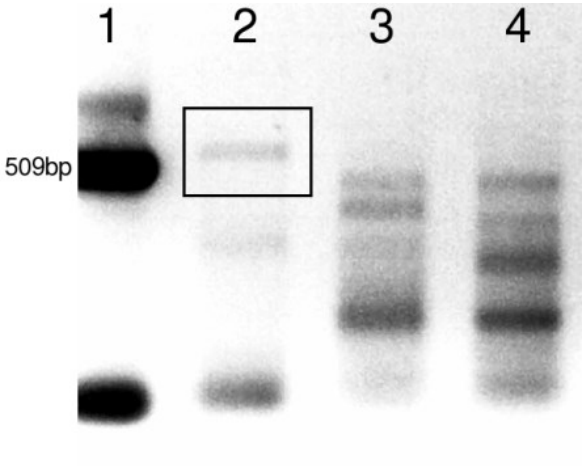


Figure 13 : Gel electrophoresis of redundant sodium channel PCR products
Lane 1 contains the positive control, ferritin like gene, which has a size of 509bp. Lane 2 includes a PCR Band from PCR with Primers 1 & 2 which has been boxed and appears to be the correct size based on predictions using homologous genes. This was gel extracted, cloned and sequenced and returned a partial sodium fragment which aligned with the ~550bp region in homologues genes. Lanes 3 and 4 are PCR products from PCR with Primers 3 &4 and Primer 5 & 6 with predicted protein sizes of 1.4Kb and 1.5kb respectively. These amplification reactions were unsuccessful in producing products of large enough size.

down PCR products with combinations of degenerate primers. PCR products were cloned and sequenced as previously described.

Figure 14 and Figure 15 show the cDNA and protein sequences of two fragments encoding part of putative *A. californica* sodium channel genes. Two 5' gene specific primers (GSPs), Table 9, were designed for each of the sequences produced and degenerate primers 7-9 designed. These 5' GSP and 3' degenerate primers were assessed for products in order to try and amplify larger fragments which included unknown 3' regions. Despite effort to retrieve more of the 3' region

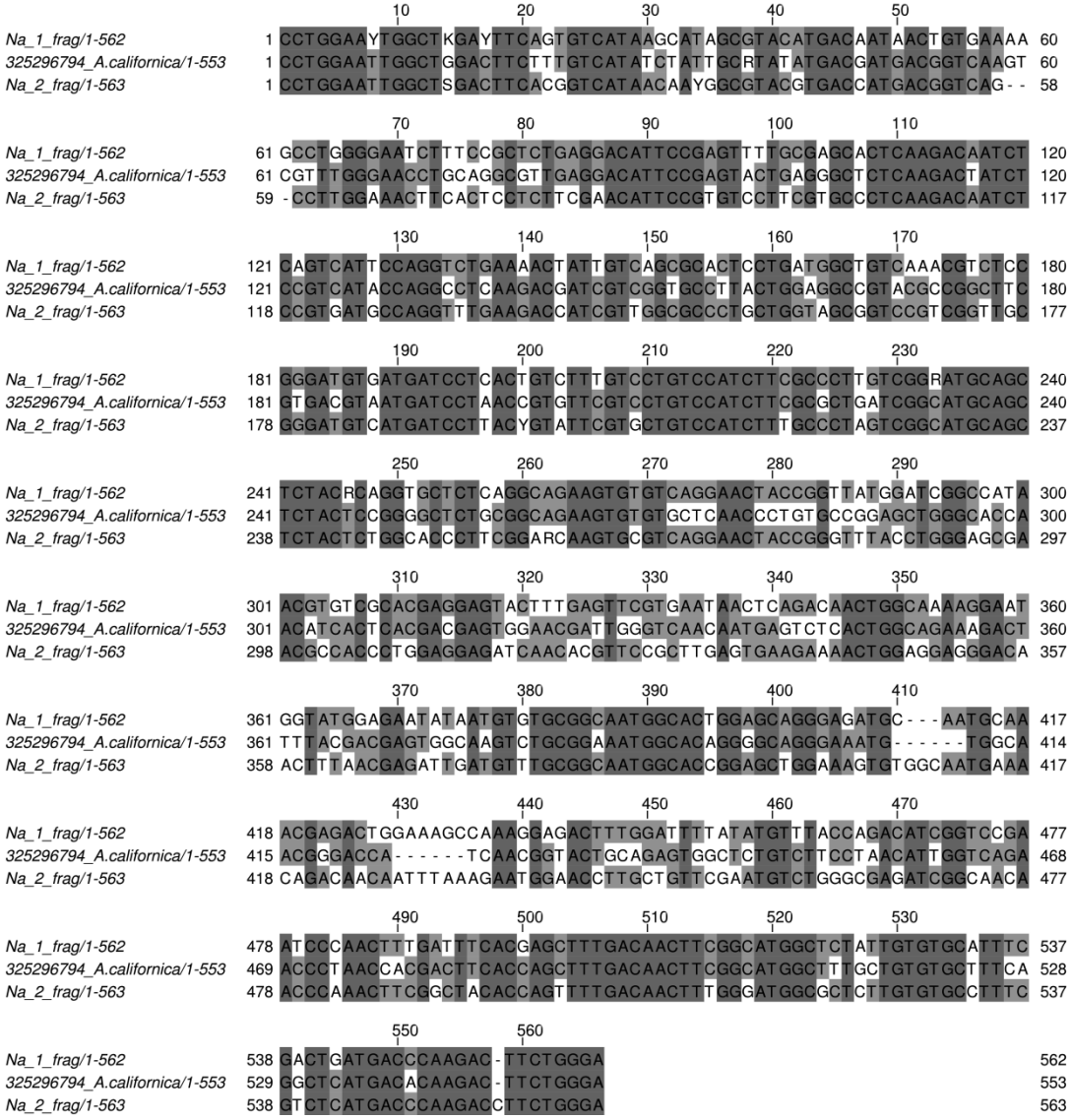


Figure 14 : Alignment of D. reticulatum cDNA sequence with A. californica SCAP1 gene

2 cDNA sequences synthesised from Deroceras reticulatum show strong homology to the 910bp -1463bp region with the 6396bp A. californica SCAP1 gene (GI: 325296794). This region represents the region between primers 1 and 2 indicated in previous figures.

of the gene, the length of sequence caused technical difficulties. Random hexamers which were used to produce the cDNA, improves the chances of 5' regions of RNA being reverse transcribed, but often decreases average cDNA length.

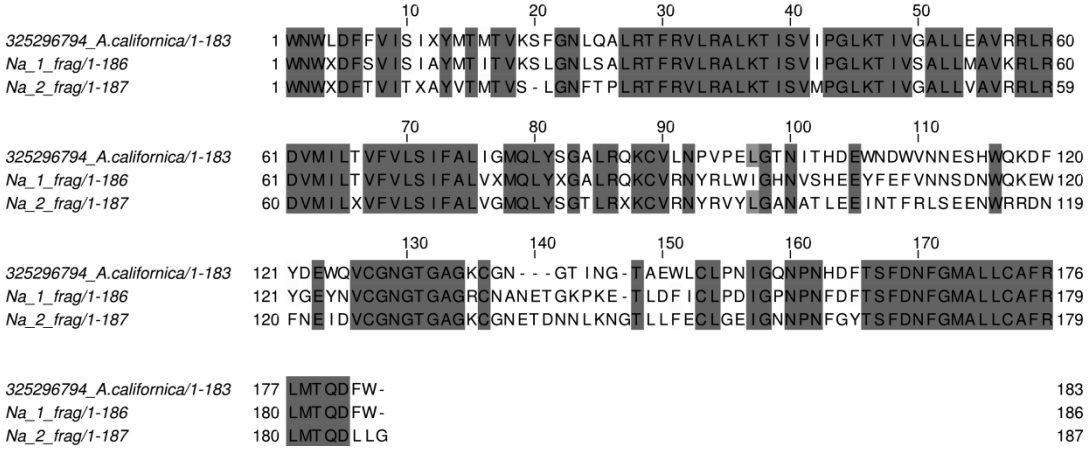


Figure 15 : Alignment of D. reticulatum protein sequence with A. californica SCAP1 gene

Alignment of translated D. reticulatum sodium channel fragments and translated A. californica SCAP1 gene. Both D. reticulatum and A. californica sequences contain ambiguous bases, most likely due to the low level of transcription leading to the high level of polymerase amplification required. Several amino acid additions are seen in the D. reticulatum gene, including 3 residues at around 140aa and 1 at 147aa which is only seen in 1 of the fragments.

Name	Sequence	Position (bp)	Fragment
5' Na_1 VNN	AGTTCGTGAATAACTCAGACAACTGG	1241	Frag 1
3' Na_1 MALL	AAATGCACACAATAGAGCCATGCC	1453	Frag 1
5' Na_2 SEEN	GTTCCGCTTGAGTGAAGAAAACTGG	1239	Frag 2
3' Na_2 CAF	ATGAGACGAAAGGCACACAAGAGC	1461	Frag 2

Table 9: Table of Gene specific primers for D. reticulatum Sodium Channel Fragments

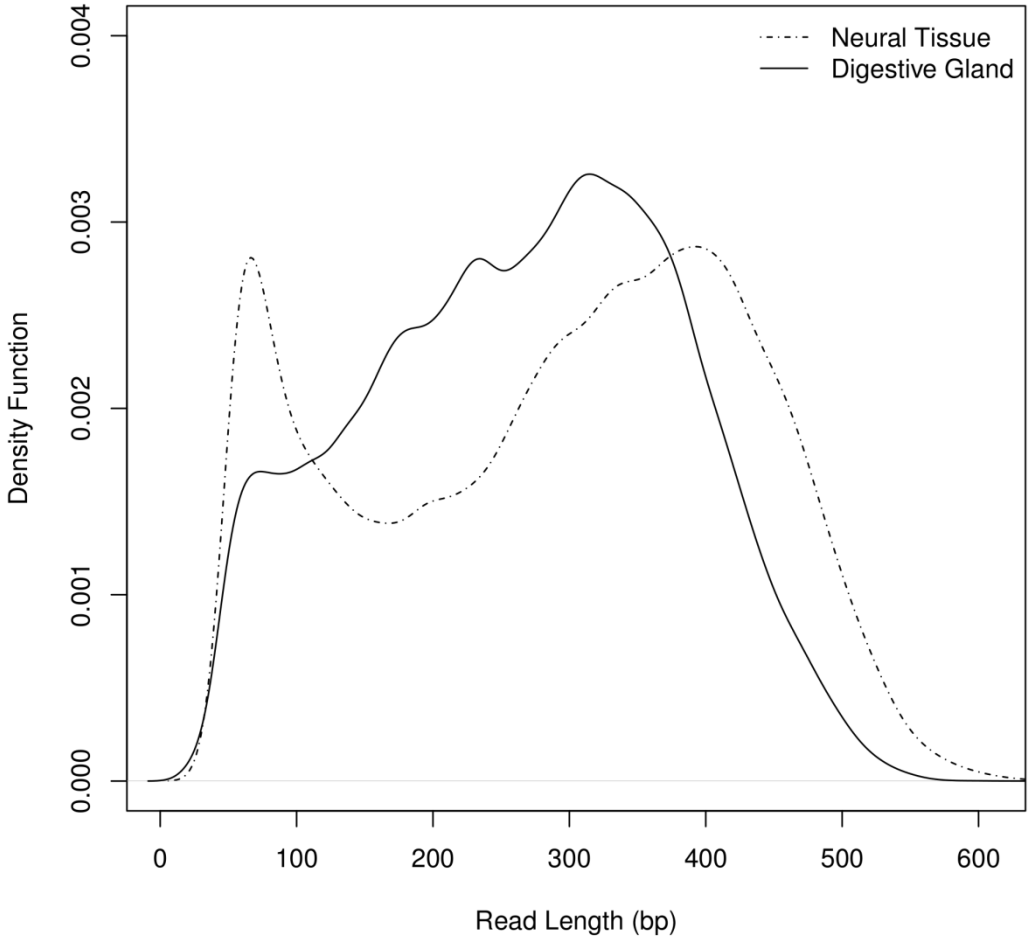


Figure 16 : Kernel Density plot of 454 reads for Neural and Digestive tissue sources

Similar to a histogram, a kernel density plot represents distribution, in this case of read length, within a dataset. The digestive gland presents an expected distribution peaking at around 300bp. The neural tissues main distribution has slightly lower density, though is shifted toward larger reads. It includes an additional peak, not seen with the digestive gland at around 75bp.

3.4 High Throughput Sequencing

3.4.1 454 Sequencing

Having evaluated the cDNAs prepared from *D. reticulatum* tissues, high-throughput sequencing was undertaken. Non-normalised cDNA from *D. reticulatum* was synthesised from RNA extracted from the digestive gland and checked on gel, see Figure 16, and with the aforementioned random sequencing. cDNA was sequenced using a 454 GS FLX platform by The Food & Environment Agency (FERA) (Sand Hutton, York, YO41 1LZ) in July 2010. This resulted in the production of 132,597 reads with an average read length of 265bp; a kernel density plot of read length is shown in Figure 16. Non-normalised cDNA from *D. reticulatum* was synthesised from neuronal tissue total RNA with an additional mRNA purification step using the Poly(A)Purist™ Kit [Life Technologies]. The resulting cDNA was sequenced at FERA in January 2012. This produced 161,419 reads with an average read length of 285bp, the distribution of which is also shown in Figure 16.

3.4.2 Sequence Cleaning with Univec

Digestive gland tissue SMART primer sequences were removed by FERA as part of their workflow, for neuronal tissue SMART primers were not removed. Adaptors and primers were removed from both 454 datasets by screening against the Univec NCBI database (which contains common sequencing adaptors). Vector screening was conducting locally using BLASTn, as the online tool did not permit such large datasets, with parameter specification matching VecScreen. The resulting BLAST output was assessed using the criteria specified by NCBI with strong, moderate & weak matches depending on BLAST score and terminal proximity. All sequences with a weak match or higher were trimmed removing the matching region. In the case of internal areas the sequence was split into two sections. Any sequences that were subsequently shorter than 50bp were removed completely. Figure 16 shows a summary of the trimmed sequences for both sequencing runs. For digestive gland tissue >95% of matches were GS FLX Titanium adaptors (Univec accessions: 929 & 933). Neuronal tissue matched various cDNA related adaptors primarily from the

SMART kit, additionally general purpose adaptors were found, which were possibly

	Digestive Tissue	Neuronal Tissue
Starting Read	132,597	161,419
Univec matches	13,368	85,411
Right Terminal Trim	12,339	7,616
Left Terminal Trim	31	71,578
Internal Trim	79	501
Total Sequences Removed	843	5,329
Total Bps Removed	303,103	940,923
Final Read Count	131,754	156,090
Dataset Contraction	0.86%	2.03%

Table 10: Table showing statistics for Univec screen

carry-over of oligo-dT from the mRNA purification step. Although there is noticeable difference in number of removed read sequences, this is likely due to the larger number of small <100bp in the neuronal tissue dataset as seen in Figure 16. The difference in number of sequences trimmed between the two datasets is primarily due to SMART primers being removed from the digestive gland before the Univec screen by FERA, the results of which are unavailable.

3.4.3 Assembling Digestive Gland Data

454 sequencing data of digestive gland was assembled with 3 assemblers; see statistics in Table 11. Contigs smaller than 200bp were removed, either through defining the option in the assembly software, or filtered out after assembly. This was done to make assemblies equivalent, with assemblers having differing minimum contig length, 200bp was selected due it being the minimum contig length accepted by the NCBI TSA database. The statistics present quite differing results, with over 3 times as many contigs in CLCbio assembly as Newbler. The n50 statistic is largest for Newbler; this statistic represents the fewest number of contigs which can represent 50% of the assembly. Despite this the Newbler assembly contains far fewer contigs larger than 1kb in comparison to CLCbio. This may suggest some of the very

Statistic	Newbler	SeqMan NG	CLCbio
No. of Contigs	1302	2570	4614
Assembly Size	1,054,550bp	1,402,125bp	2,606,554bp
n50	**950**	621	613
n90	**431**	284	325
Contigs >500bp	861	766	**1958**
Contigs >1Kbp	298	272	**413**
No of Reads (%)	90973 (69%)	98849 (75%)	**110830 (84.1%)**
No of bps (%)	24,323,911 (69.7%)	24,968,338 (71.5%)	**28,0157,27 (80.2%)**
Max Contig (bp)	4437	**4527**	3768

Table 11: Statistics of 3 Assemblies of Digestive Gland 454 Read Sequences

Reads were assembled with Newbler v2.6, SeqMan NG v4.1.2 and CLCbio 4.7.2. Emboldened statistics are where metrics can be considered superior with CLCbio having more contigs larger than 0.5 & 1Kb and incorporating more reads/bps. Newbler has better n50 & n90 statistics.

large contigs have fragmented into mid-sized contigs in CLCbio, or vice-versa Newbler has over-assembled mid-sized contigs into chimeric contigs. Notably, the assembler which overall incorporated the most reads is CLCbio with Newbler being poorest. Whilst SeqMan NG appears to sit in between Newbler and CLCbio, when we look at the distribution of contig size, SeqMan NG has greatest skew to shorter contigs, see Figure 16. Overall both CLCbio and Newbler have advantages based on different metrics, whilst there seems no reason to prefer the SeqMan NG assembly.

3.4.4 Assembling Neuronal Tissue Data

454 neuronal tissue data was assembled with 3 assemblers as was done for the digestive gland data, statistics are shown in Table 12. Overall assembly statistics appear relatively varied although Newbler has the best metrics for a number of statistics such as n50, contigs larger than 1kb as well as incorporating the most base

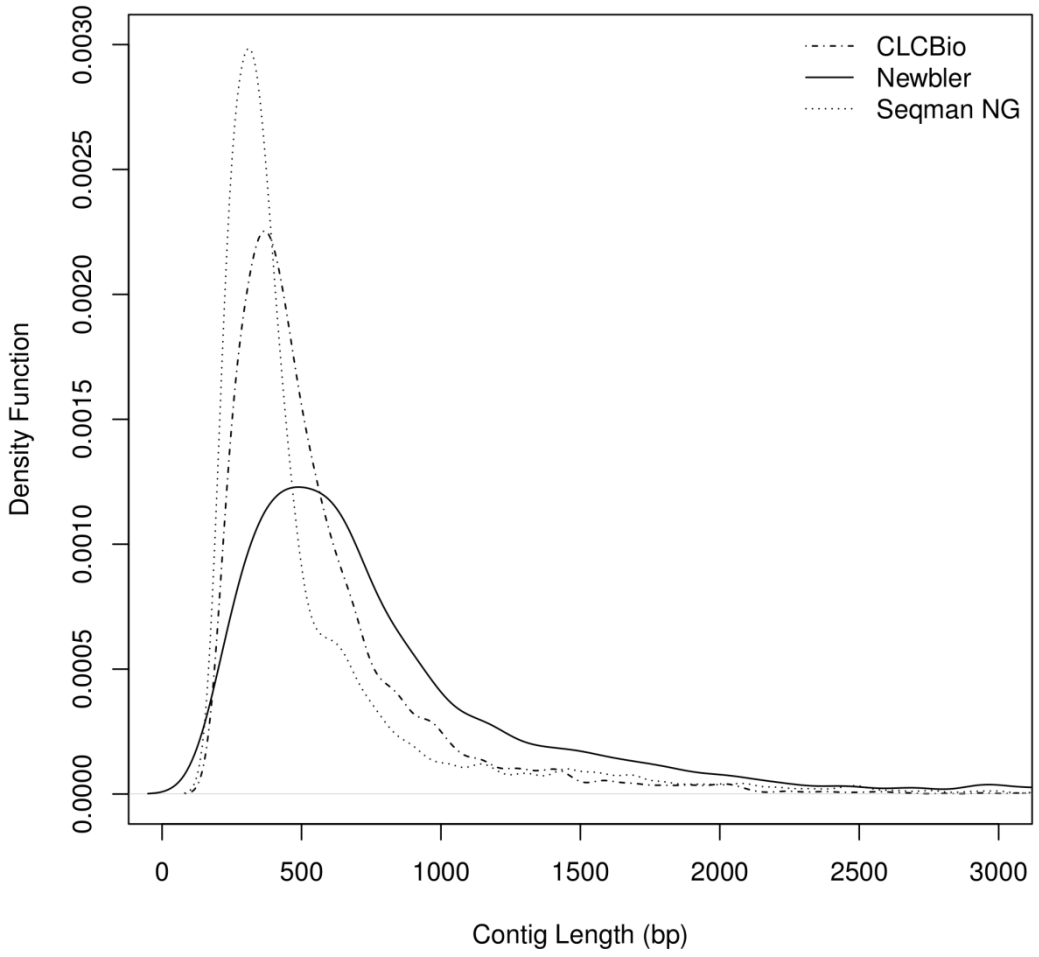


Figure 17 : Kernel density plot of digestive gland contig lengths

The graph shows the density of contigs length relative to the overall assembly size. Overall Seaman NG is skewed further toward small contigs with a much higher peak. The Newbler assembly presents a much lower kurtosis distribution, with the CLCbio in an intermediate position. It would appear Newbler would be the closest assembly to more optimal standard distribution, though may not represent the true distribution any more than the other assemblies.

Statistic	CLCbio	Newbler	SeqMan NG
No. of Contigs	8018	2614	6509
Assembly Size (bp)	3,959,511	1,736,684	3,158,950
n50	527	**745**	502
n90	312	**397**	306
Contigs >500bp	**3040**	1627	2097
Contigs >1Kbp	302	**393**	304
No of Reads (%)	**118021 (75.6%)**	98556 (63.1%)	92976 (59.6%)
No of bps (%)	32254579 (71.4%)	**3,2337,120 (71.5%)**	25774512 (57%)
Max Contig (bp)	2130	2476	**2490**

Table 12: Statistics of 3 Assemblies of Neuronal Tissue 454 Read Sequences

Reads were assembled with Newbler v2.6, SeqMan NG v4.1.2 and CLCbio 4.7.2. Emboldened statistics are where metrics can be considered superior. Overall the size distribution related statistics such as n90 and contigs >1Kbp are very similar for all assemblers. Overall SeqMan appears to be poorest, including least number of base pairs and poorest size statistics.

pairs. The maximum contig size is much lower than that of the digestive gland. The neuronal tissue cDNA was prepared with the additional mRNA enrichment step, this difference may account for the difference in contig size. Alternatively, as shown in the cDNA gel analysis, see Figure 9, the additional time taken may have led to poorer RNA resulting in a smaller size distribution of cDNA sequences. However comparison of contig size using a kernel density plot shows only a very minor difference between the two datasets size distribution, Figure 17. The overall density is shifted slightly further toward smaller contigs in the neuronal dataset, suggesting overall at slightly poorer quality dataset which is mirrored in the overall statistics seen in Table 12. The table shows more contigs, but overall shorter in length. Despite this we see very similar levels of incorporation of reads into assemblies.

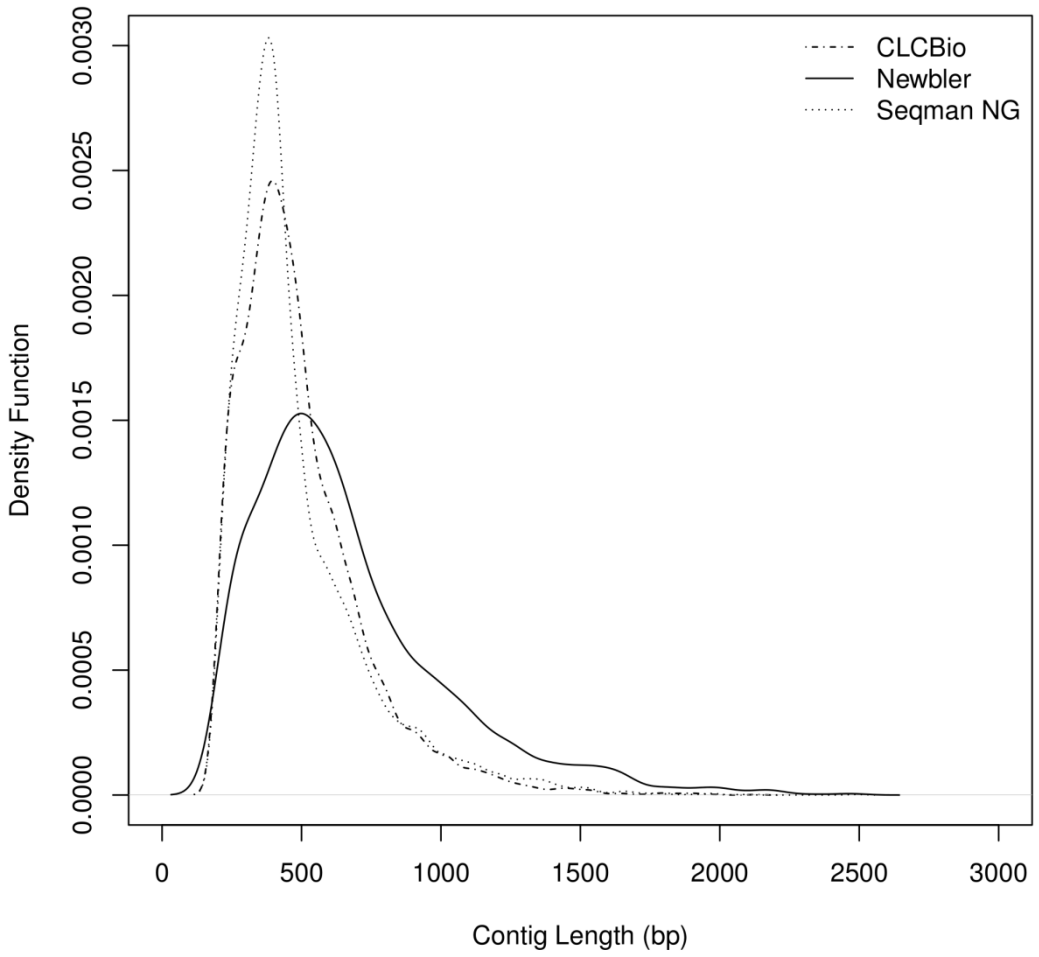


Figure 18 : Kernel density plot of neural tissue contig lengths

The graph shows the density of contigs length relative to the overall assembly size. In comparison to the digestive gland contigs, the peak in density of distribution is slightly closer between assemblies though the peak density at slightly shorter contigs and the overall extend of the density toward large contigs is contracted as compared with the digestive gland assemblies.

3.4.5 Comparison of Assemblers

The ratio of results did not match previous data comparison conducted which included these assemblers (Kumar and Blaxter 2010), but assemblers used here are

later versions, and the datasets used were roughly 1/5th the size. Notably the ratio of number of contigs 1:2.5:3 in the neuronal tissue assembly follow the same rank as with the digestive gland with a ratio of 1:2:3.5, all ratios being ranked Newbler, SeqMan, and CLCbio in number of contigs. There are a number of discrepancies in the actual size distribution and inclusion rates of the assemblies. The CLCbio digestive gland assembly appears to use the largest amount of reads to assemble the most contigs, with the largest number of contigs greater than 500bp and 1Kbp. Although Newbler does produce the best n50 of the assemblies this equally could be due to fewer smaller contigs or CLCbio under-assembly of larger contigs. Overall comparison through basic assembly methods is relatively uninformative; utilization of homology data whilst having its caveats, should be more informative.

3.4.6 Contig Naming

In order to represent contigs which exist in multiple assemblies, custom naming convention has been used for the contigs. This nomenclature is best attempt to explain the relationship of a sequence with respect to multiple source assemblies. CLCbio contigs have prefix C then number referring the position in the output assembly; Newbler has prefix N and SeqMan prefix S. Where contigs are equivalent or within a 90% threshold of similarity defined by BLAST search, and their equivalence checked by hand, against the assembly there are concatenated together for example C1212;N3233;S2332n, with a terminating suffix referring to either the digestive or neuronal assemblies (d or n). In the case of contigs matching but being incomplete prefix p is used and - is for the contig missing from the assembly entirely. For example C1212;N-;pS2332n would refer to CLCbio contig 1212 which does not exist in the Newbler assembly and is only partially assembled in the SeqMan assembly. On one occasion a second contig from digestive gland is added in parenthesis to a neural contig for example C-;pN48;S-n(C2550;S1758;N554d) which is used to identify a partial contig in the neural tissue dataset which exists in a complete form in the digestive gland dataset.

Chapter 4 | Analysis: Transcript Sequences

4.1 BLASTx homology analysis of Digestive Gland transcriptome data

The predicted transcripts from cDNA sequencing (both contigs and singletons) were analysed using BLASTx, which compares the protein sequences predicted by transcript sequences with protein sequences in the NCBI non-redundant (nr) protein database [last conducted March 2013]. Searches against protein sequences rather than nucleotide searches are preferred as this both reduces the search time and the resulting data is more informative. Transcript to DNA homology being less informative than transcript to protein homology, and allows for more accurate functional predictions. The significance of any sequence similarity detected is measured by the e-value parameter attached to each comparison, which estimates the probability of the similarity being due to random chance. Data was uploaded to a modified BioSQL database which allowed for a wide range of statistical analyses. Altering the BLAST e-value “cut-off” had a significant effect on the number of sequences detected as significantly similar to the predicted *D. reticulatum* protein sequences, and using too low a "cut-off" value removed similarities which appeared significant on manual inspection. This is likely to be due to the smaller number of closely related species to *D. reticulatum* with sequences in the nr database, which meant that the effective database size was smaller than the actual database size. Figure 19 shows how changing e-value affects the number of contigs which match homologues for each of the assemblies with Table 13 showing a summary at 1e-3; this value is higher than that normally used as a "cut-off" for sequence similarity comparisons, but was used as an empirical compromise between detection of too many "background" similarities, and omission of too many relevant similarities.

The analysis of sequence similarities was used to generate the parameter "number of reads with hits". This value is a summation of the total reads that make up each of the contigs which have a BLAST hit. Table 13 shows that despite the disparity in BLAST matches for contigs produced by each of the assembler softwares

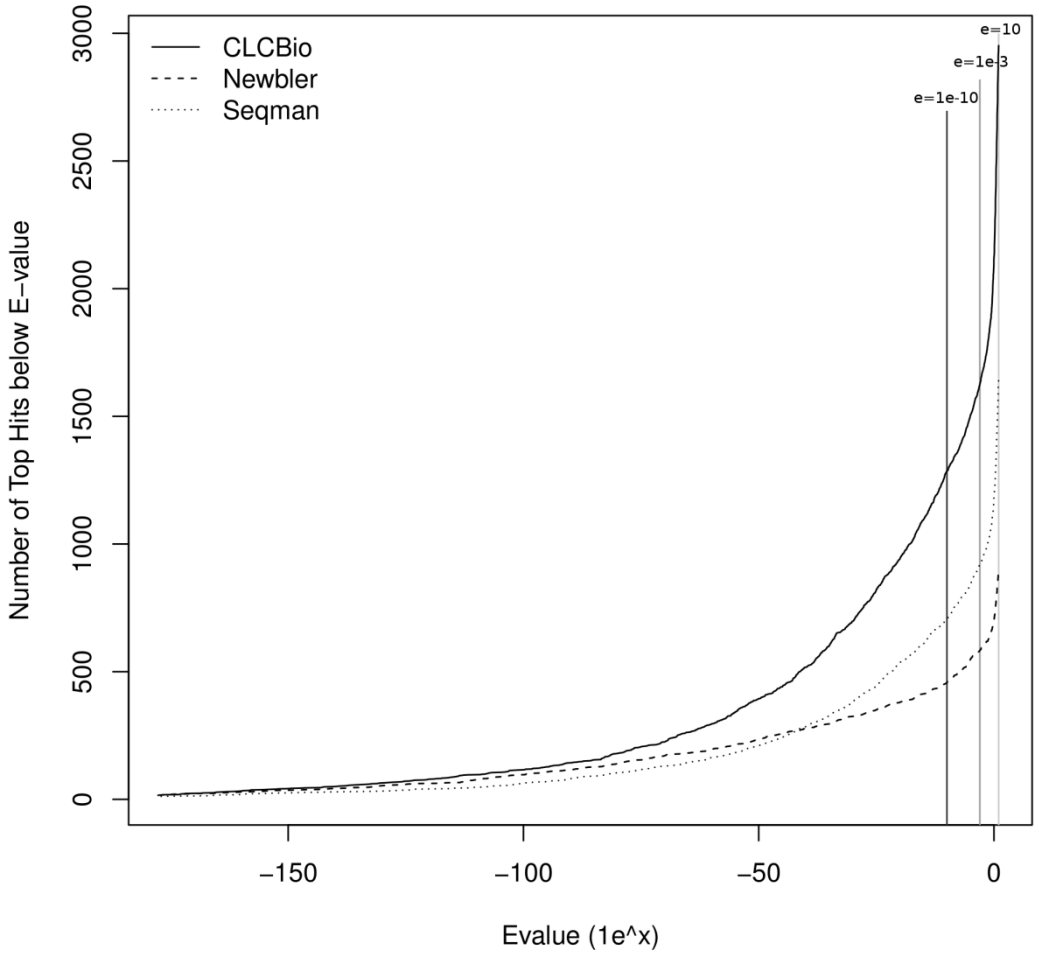


Figure 19 : Number of top BLAST hits below BLAST evalue for 3 assemblies

The graph represents the change in number of contig matches for digestive gland BLAST results which would be filtered out of the results depending on the e-value cut-off. Expect values of 1e-10, 1e-3 and 10 are shown as vertical lines with the former 2 representing common cut-offs and the later representing the limit of expect value for a BLAST search.

Summary (All expect < 1e-3)	CLCbio	Newbler	SeqMan NG
Contigs with BLAST matches	1626	583	921
Total NCBI protein homologues	55634	29796	44483
Number of reads with hits	31922	30129	29215

Table 13: Summary of BLAST homology matches for digestive gland assemblies

The contigs with BLAST matches describes the number of contigs which have at least 1 match to a protein sequence in the NCBI database which has an evalue lower than 1e-3 and matches the 1e-3 line of Figure 19. The total NCBI protein homologues is the number of unique proteins that match a contig, these are only proteins which match at expect value below 1e-3 and are included in the top 50 hits of a BLAST result.

used for raw data analysis, the overall number of reads which have BLAST matches is very similar between the assemblies. However, this is a raw count and does not indicate whether the ~30,000 reads which have BLAST hits in one assembly are the same 30,000 in another. However, using BioSQL it is relatively simple to count how many reads overlap between assemblies; Figure 20 represents this data. Roughly 80% of reads from each assembly which contribute to contigs with sequence similarity to the global database ("BLAST hits") correspond to reads from at least 1 other assembly which satisfy the same criterion. This leaves 20% of reads from each assembly constituting "BLAST hits" unique to that assembly. For this reason, it was beneficial to continue to use data from all assemblies, rather than choosing one assembly and not considering data from the others.

4.1.1 Species and Phyla of homology matches

At the time of analysis, the NCBI databases included only 125,529 proteins for Mollusca, compared with 3,582,193 and 1,788,814 for Chordata and Arthropoda respectively. These 2 phyla contain a greater variety of potential homologues to *D.*

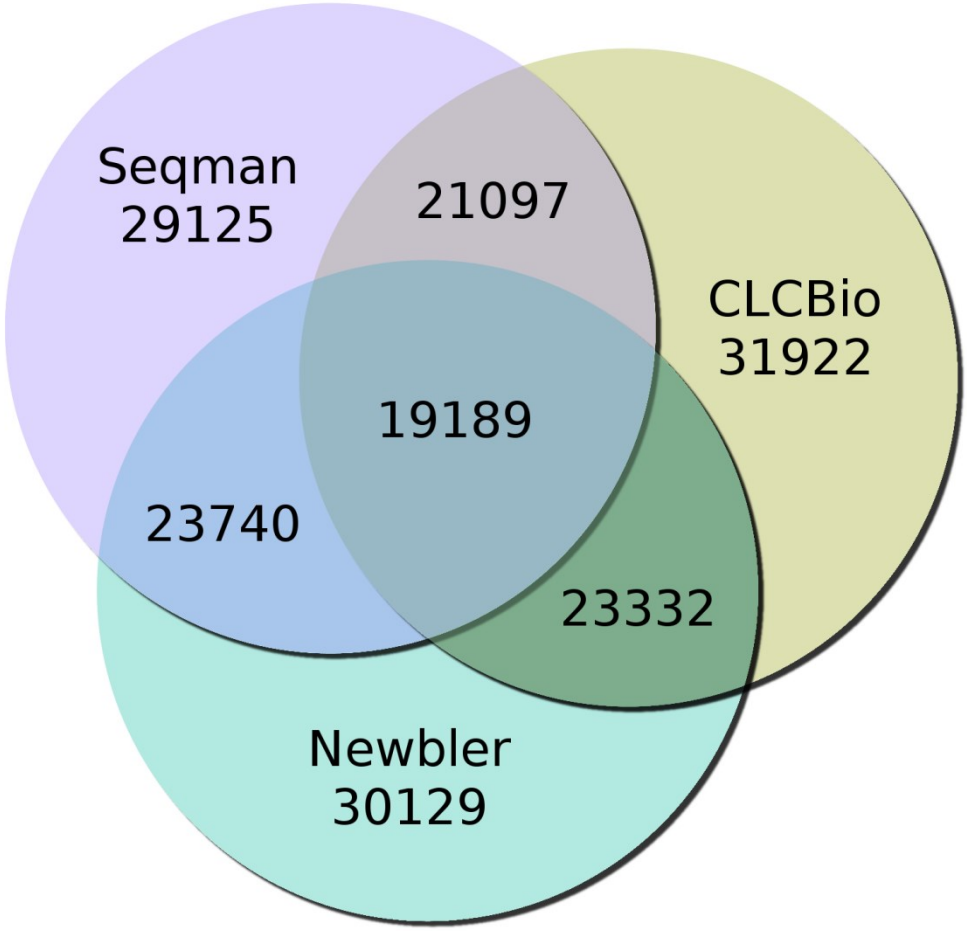


Figure 20 : Venn diagram of reads with BLAST matches below 1e-3

The Venn diagram shows how reads which are part of contigs, which have BLAST matches, are shared between assemblies. In this case 19189 reads match homologues and are in all 3 assemblies. Roughly 20% of reads per assembly uniquely match homologues and are either not included in other assemblies, or are included but have not been found to match a homologous protein.

reticulatum predicted proteins than the poorer characterised Mollusca phylum, to which *D. reticulatum* belongs, Figure 21. Despite the relative lack of Molluscan data in the NCBI non-redundant protein sequence database, approximately a third of the most similar proteins were Molluscan homologues, compared to Chordata and Arthropoda which were respectively the next largest phyla. These data suggest that

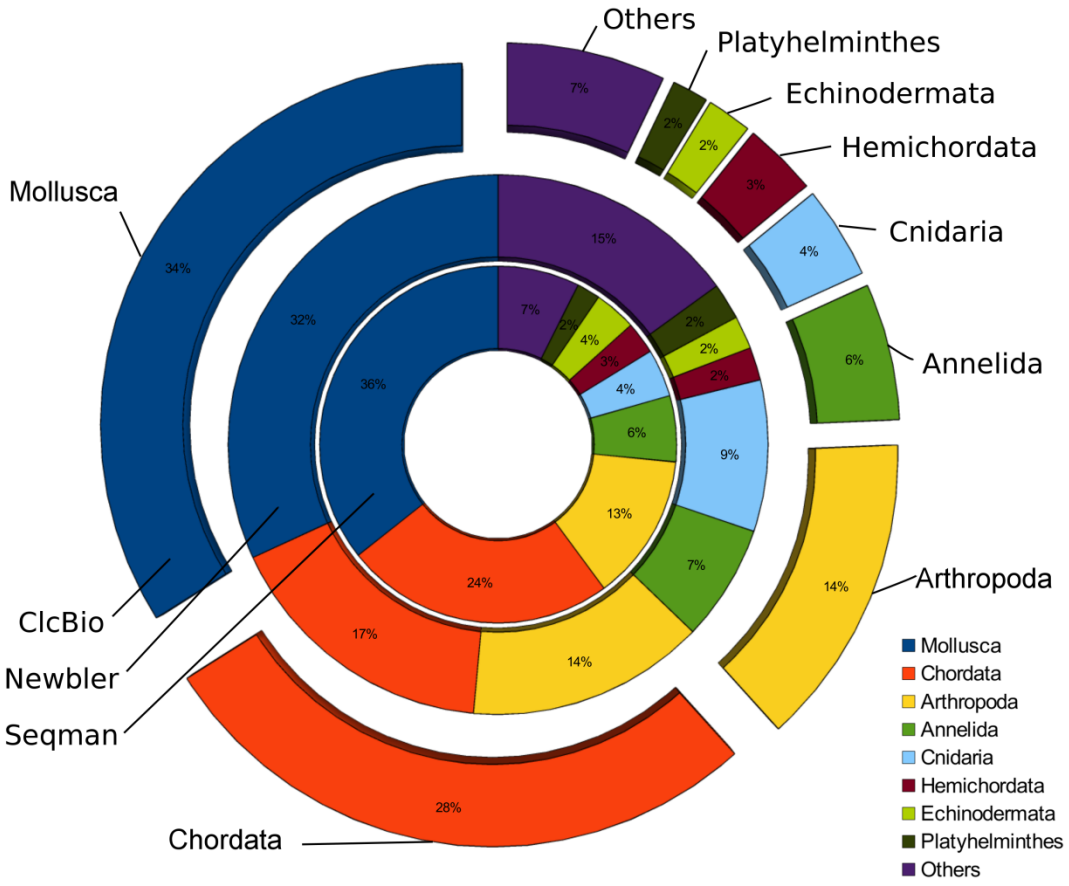


Figure 21 : Top BLAST matches per phyla as a percentage for the digestive gland

The chart shows the breakdown for the top BLAST hit per contig which had a BLAST hit with an expect value below 1e-3. The inner, middle and outer rings represent the SeqMan NG, Newbler and CLCbio assemblies. The top hits are against Molluscan homologues as one would hope for a Molluscan transcriptome. Chordata and Arthropoda we would also expect based on the distribution of the NCBI database, with these phyla having the widest range and variety of homologues.

where Molluscan homologues of the predicted *D. reticulatum* proteins are present in the database, the similarity comparison selects them.

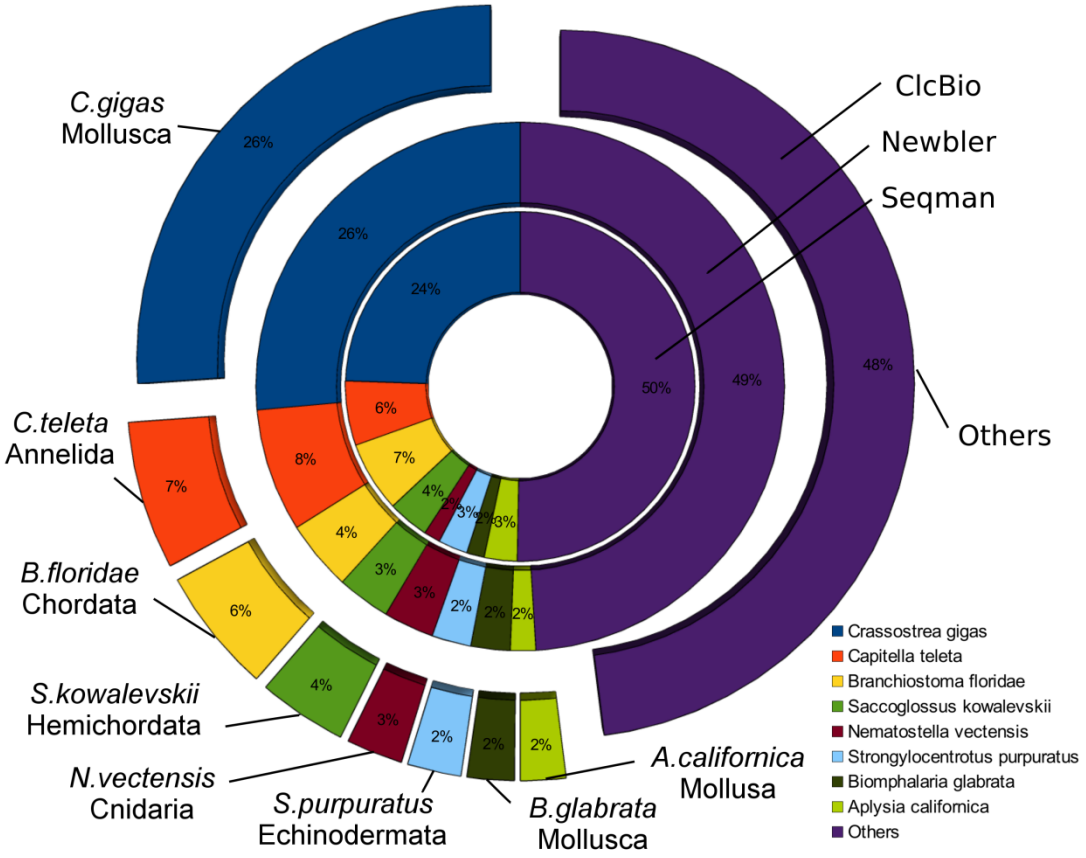


Figure 22 : Breakdown of top hits by species for the digestive gland assemblies

The chart represents the species of the protein which was the number hit for each contig in an assembly. The hits were limited to having e value of below 1e-3. The three rings of this pie chart represent the three assemblies with the outer, middle, and centre rings representing CLCbio, Newbler and SeqMan assemblies respectively. We see in general that the species with biggest match is the only Molluscan species which has a genome in the NCBI database, as would be expected. Other species mostly represent the next most characterised organism in each respective phylum.

The top species returned from similarity comparison, with approximately a quarter of most similar proteins, was *Crassostrea gigas*, Figure 22. This is as we would expect, with *C. gigas* being the only mollusc to have a complete genome

sequence with all predicted proteins added to the NCBI nr protein database. The fast-developing nature of the global sequence database was dramatically illustrated by the change in sequence comparison statistics which took place during the duration of this project.

Notably, when the digestive gland sequences were first analysed for sequence similarity to database entries in 2010 the top species detected was *Branchiostoma floridae*, a Lancelet of the Chordata, which gave 20% of most similar "BLAST hits" in 2010; reanalysis of the data in 2013 showed this species having <7% of the most similar "BLAST hits". There were also are several other Chordata species, *Xenopus tropicalis* and *Danio rerio*, both with ~4% of top hits in 2010, which are now both <1%. The decrease in Chordata species and increase in Molluscan and other invertebrate matches indicates an increasing number of less characterised phyla are now being better represented in the NCBI database. This is shown in BLAST analysis on a *Lymnaea stagnalis* transcriptome from 2012, using refseq and swissprot NCBI databases from between November 2011 and February 2012 which showed Molluscan homologues representing only 1.6% of the top hits (Sadamoto et al. 2012).

4.2 ESTScan peptide prediction and InterProScan

Predicted polypeptides for assembled sequences were produced using ESTScan2.1 (Iseli et al., 1999) to identify likely ORFs. The resulting predicted polypeptides were then analysed with InterProScan, following a bioinformatic workflow previously described (Cantacessi et al. 2010). Whilst this method increases the chances of sequences being missed and not analysed, it substantially decreases computation time if nucleotide sequences are first translated, leading to 5x more sequence comparisons to the proteins. InterProScan is a collection which uses 14 separate software packages to analyse the peptide sequence for peptide motifs and homologous domains (Zdobnov and Apweiler 2001). This package includes software which compares to known protein databases such as PANTHER, and protein families and motifs as well as amino acid signatures such as in the PROSITE database. Additionally software such as SEG, COILS and SignalP detect specific amino acid

Main Term	IPR Group	Average	CLCbio	Newbler	SeqMan
IPR013128	Peptidase C1A, papain	**1871**	1972	1864	1776
IPR001701	Glycoside hydrolase, family 9	**1374**	1852	1984	287
IPR009078	Ferritin	**1139**	1085	1161	1170
IPR020568	Ribosomal protein S5	**750**	714	813	723
IPR000217	Tubulin	**663**	680	638	670
IPR005225	Small GTP-binding protein domain	**499**	334	589	573
IPR001404	Heat shock protein Hsp90	**444**	447	448	438
IPR014001	Helicase	**424**	433	436	404
IPR001299	Ependymin	**398**	401	405	388
IPR012336	Thioredoxin-like fold	**396**	424	409	356
IPR003594	Histidine kinase-like ATPase	**392**	395	395	387
IPR003726	Homocysteine S-methyltransferase	**378**	346	394	394
IPR015566	Endoplasmin	**377**	377	383	372
IPR016186	C-type lectin-like	**343**	360	342	326
IPR001753	Crotonase superfamily	**342**	155	447	425
IPR000640	Translation elongation factor EFG	**334**	261	377	365

Table 14: Summary of top IPR terms

The table shows a list of the top IPR terms ranked by the number of reads linked to the IPR terms through a contig match using the IPRscan software. IPR terms have been condensed into groups and the IPR term with the highest number of read matches used. IPR terms linked to generic domains such as EF Hand domain, NAD(P)-binding domain, Leucine-rich repeat et cetera have been removed as they do not indicate any likely function and as such are not relevant. Additionally major IPR terms which are only significant in one of the assemblies have not been included, but are shown in a separate table.

configurations linked to protein structures such as coiled-coils and signal peptides rather than homology. Specific matches made by these softwares are then linked to IPR terms, which in turn can be linked to protein families as well as gene ontologies.

IPR Term	IPR Name	CLCbio	Newbler	SeqMan
IPR011687	P60-like	**1334**	10	5
IPR005485	Ribosomal protein L5 eukaryotic/L18 archaeal	**1016**	12	6
IPR008983	Tumour necrosis factor	**12**	8	518
IPR001073	Complement C1q protein	**4**	0	514
IPR000163	Prohibitin	**33**	328	24

Table 15: Summary of IPR terms only significant in a single assembly

The table shows IPR terms which were not included in the top IPR term summary table as they were not represented highly in all the sequence assemblies. With the sequences having both been detected as a protein with ESTScan and matched to known motifs with IPRscan, it is likely these are real proteins. However there significance with high read count may be an artefact of assembly, possible caused by over assembly of multiple contigs.

4.2.1 Analysis of IPR terms for Digestive Gland tissue

Table 14 lists the Top 20 IPR terms based on number of reads linked to the IPR term. Using the IPR terms we can predict potential functions for the sequences and using read counts over simply number of contigs both allows for ranking to better represent the availability of transcripts in the sequencing results as well minimising the effect of assembly quality. As Table 14 demonstrates, whilst there are differences, in general the same main proteins are represented by the data. Where the assemblies uniquely match IPR terms these have not been included, but are shown in Table 15. The top proteins represent a mix of structural, storage and enzymatic enzymes. Notably, peptidase C1A, papain-like cysteine protease, is the top IPR group, which is not unexpected for digestive gland tissue and matches previous

IPR Term	Active Site Name	Average	CLCbio	Newbler	SeqMan
IPR025660	Cysteine peptidase, histidine	**1791**	1800	1816	1757
IPR000169	Cysteine peptidase, cysteine	**1756**	1788	1802	1677
IPR025661	Cysteine peptidase, asparagine	**1685**	1859	1857	1338
IPR018221	Glycoside hydrolase, family 9	**1314**	1707	1958	277
IPR020830	Glyceraldehyde 3-phosphate dehydrogenase	**335**	293	394	319
IPR018114	Peptidase S1, trypsin family	**268**	274	266	265
IPR020003	ATPase, alpha/beta subunit, nucleotide-binding	**154**	196	145	122
IPR011767	Glutaredoxin	**154**	151	157	154
IPR022415	ATP:guanido phosphotransferase	**113**	101	129	108
IPR020610	Thiolase	**109**	85	126	117
IPR020615	Thiolase, acyl-enzyme intermediate	**103**	85	116	109
IPR008271	Serine/threonine-protein kinase	**100**	69	119	111
IPR019826	Carboxylesterase type B	**77**	155	47	28
IPR024708	Catalase	**57**	9	154	8
IPR018220	Adenylosuccinate synthase	**57**	59	54	57
IPR001252	Malate	**25**	62	13	0

	dehydrogenase				
IPR001969	Peptidase aspartic	**8**	12	11	0
IPR001579	Glycoside hydrolase chitinase	**5**	16	0	0
IPR016130	Protein-tyrosine phosphatase	**4**	7	6	0
IPR002168	Lipase, GDXG	**3**	9	0	0
IPR012999	Pyridine nucleotide-disulphide oxidoreductase, class I	**2**	6	0	0

Table 16: Top active sites based on IPR terms

Top active sites based on IPR terms labelled as active site type. This should indicate the best represented enzymes in the dataset. The top active sites represent digestive enzymes sites we would expect to see in digestive gland tissue. SeqMan's result for GHF 9 site seems anomalous and may suggest GHF9 is more significant than would appear, having most reads of any site in the Newbler assembly.

experimental evidence (Walker et al., 1998). In addition other key digestive enzymes are found in the top group, these can be further clarified by filtering only active sites, shown in Table 16. The top 3 main digestive enzymes are cysteine peptidase, chymotrypsin-like serine protease and GHF9 (cellulase) all of which match what one would predict for digestive gland source tissue. Below we consider some of the notable proteins in more detail.

4.3 Largest protein groups based on IPR terms

4.3.1 Cysteine Peptidase

A number of different sequences encoding cysteine proteinases are linked to the Peptidase C1A IPR terms. For *D. reticulatum*, all 3 assemblies produce at least 7

Contig	Active Sites Residues	Blast Homology	CLCbio Reads	Newbler Reads	SeqMan Reads
C2952;N232;S21d	3/3	Cathepsin L	1578	1567	1169
C2566;N257;S315d	3/3	Cathepsin B	79	77	75
C4050;N10;S-d	3/3	Cathepsin C	65	84	-
C4049;N279;S529d	2/3	Cathepsin F	41	41	44
C602;N73;S-d	3/3	Cathepsin L	19	26	-

Table 17: Contigs with homology to cysteine peptidase like genes

Table includes peptides with the cysteine peptidase active site IPR terms with all but 1 peptide having all 3 sites. The peptide with closest homology to Cathepsin C includes the Cathepsin C Exclusion Domain (IPR014882) and both Cathepsin L like and Cathepsin B like peptides include a propeptide.

contigs which match at least 1 active site (PS00139, PS00639 and PS00640). These Prosite motifs centre around 3 key amino acid residues, a cysteine, asparagine and histidine which form the catalytic triad required protease activity. Additionally all assemblies contain around 20-30 sequences with BLAST homology to cathepsins B, C, F, L, K and Z. 5 contigs in particular appear to be complete or near-complete cathepsin like proteins, see Table 17. All but 1 of these proteins contain 3/3 active site residues and show homology to various different cathepsin like peptides. The top predicted protein shows close homology to cathepsin L and is a full coding sequence which has one of the highest read counts for any contig with homology matches. The top 3 active sites in Table 16 and the top IPR term (IPR000668) in Table 14 are primarily due to this cathepsin L-like contig which represents ~1.5% of the assembled digestive gland dataset in reads.

Cathepsin L has been shown in a number of different roles in invertebrates (Pyati et al. 2009); although without characterisation of this protein we cannot be completely sure of its function in *D. reticulatum.* Comparison with other Molluscan cysteine proteases indicates the primary conserved sites are all maintained. Figure 23

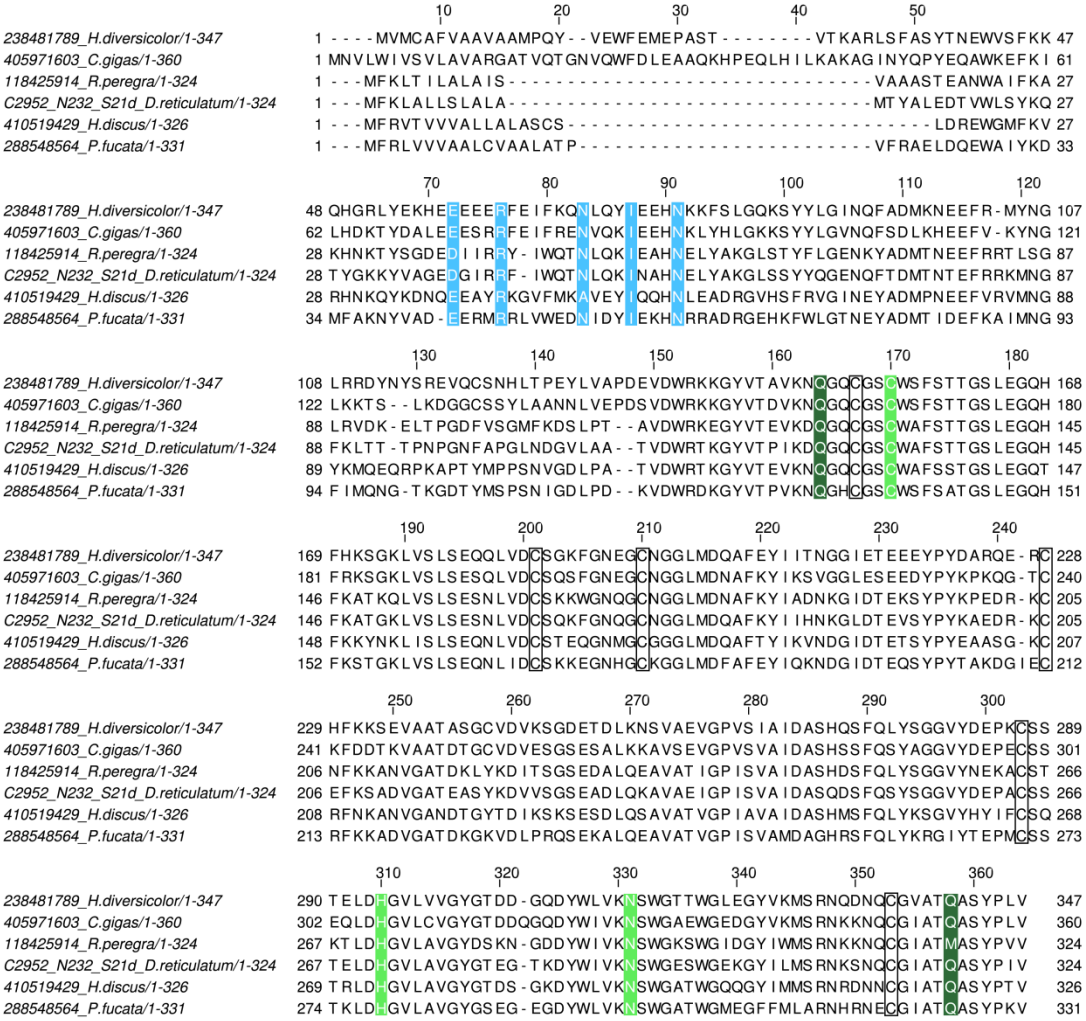


Figure 23 : Cathepsin L homologue of D. reticulatum alignment

An alignment D. reticulatum predicted Cathepsin L with Mollusca matches from the Blastx search against the nr database. Cathepsin L from Mus musculus (GI: 4887002) is also included to show that the sequence is highly conserved, even across phyla. The propeptide inhibitory domain includes an ERFNIN-like conserved region, highlighted blue, with the initial Glu and Arg and final Asn described as the most highly conserved residues across cysteine proteinases. Boxed cysteine residues in the enzymatic domain are key conserved residues which form 3 disulphide bridges as described in MEROPS database record for cathepsin L. The light green residues are the Cys-His-Asn catalytic triad essential for activity. Also highlighted in dark green is a conserved glutamine residue, at 164aa, which is essential for catalytic activity and a glutamine residue at 358aa which is indicated to be key to substrate specificity (Barrett, Rawlings, and Woessner (Eds) 1998).

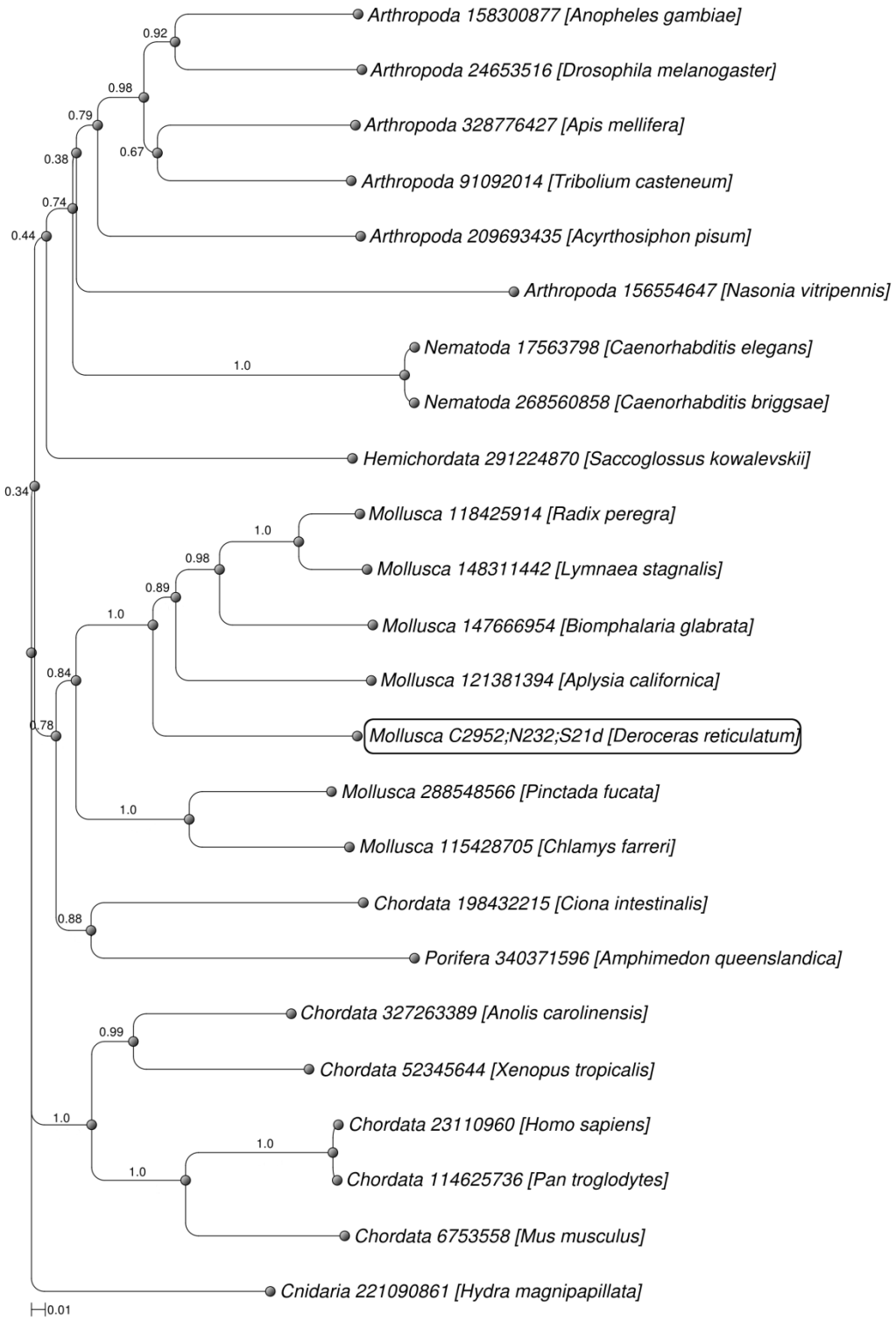


Figure 24 : Phylogenetic tree of the cathepsin L from Deroceras reticulatum

Shows a phylogenetic tree produced by aligning closest homologues identified using BLASTp from all sequences with genomes and their predicted proteins

available via NCBI Map Viewer. The tree was generated with clustalw2 using bootstrap value of 1000 (fraction shown), from 15 amino acids before the ERFNIN motif and 8 amino acids after the Asn300 conserved residue. 5' and 3' regions were left out due to missing data; many Molluscan ESTs were only partial. L. stagnalis sequence was a combined sequence of a 5' partial (148311442) and 3' partial (148311064) sequence. The contig from our data grouped with other gastropods, as part of a larger Molluscan group, with the 2 Bivalves P. fucata and C. farreri separate from the gastropods. We see that most species group into their phyla as we'd expect. A few artefacts exist primarily where only a single species is available from a phyla. These are likely due to the limitations of using a single gene and a small number of species.

shows how sequences from *Mollusca* align with *D. reticulatum*. We extended this alignment with sequences from species with full genomes, as well as some Molluscan sequences from the NCBI EST_OTHERS database to produce a phylogenetic tree, Figure 24. The Mollusca group together separate from the main Chordata and Arthropoda groups, with Arthropoda being slightly closer, though with a low bootstrap value (0.34), to Mollusca, than Chordata. A relatively high bootstrap value further separates the 2 Bivalve species from the Gastropoda.

All sequences used in the phylogenetic tree contain the key conserved sites (except *A. californica* which lacked Asn300 equivalent residue) within the peptide region as well the ERFNIN motif in the inhibitory pro-peptide.

The inhibitory domain is needed to keep the peptide in an inactive state whilst being transported to the extracellular space, where it is cleaved to become active (K. Tao et al. 1994). The IPRscan identifies a signal peptide region at the 5' end of the protein which is necessary for secretion of the enzyme for extracellular digestion. After its export to the extracellular space, cathepsin L is activated by cleavage of the pro-peptide region. Autolysis of cathepsin L at pH 6 and below has been described (Jerala et al. 1998) as well as activation by trypsin/chymotrypsin digestion (Nishimura, Furuno, and Kato 1988; Wiederanders and Kirschke 1989). A search of cDNA for trypsin shows a *D. reticulatum* contig (C4138;N153:S219d), homologous

to serine proteases. As is seen in the top active sites IPR018114 Peptidase S1, trypsin family is the next most highly represented proteinase. The InterProScan also shows a SignalP cleavage site for the serine protease suggesting it is a secretory protein, similar to insect digestive cathepsin L-like enzymes, see Figure 25.

As well as homology, biochemical characterisation of cysteine proteinases with cathepsin-L like activity was shown to be responsible for the majority of the digestive gland's proteolytic activity in *D. reticulatum* against a protein substrate (plant Rubisco; (A. J. Walker, Glen, and Shewry 1998)). These data also highlighted a lack of aspartic peptidase activity in *D. reticulatum* digestive gland due to pepstatin not reducing proteolysis of azocasein. A search through the data for aspartic peptidases showed several contigs linked to relevant IPR terms IPR001461 and BLAST hits against invertebrate aspartic peptidase / cathepsin D. The average number of reads linked to aspartic peptidase (IPR001461) is 8 (see Table 16) suggesting a limited number of transcripts in the RNA, correlating with the lack of activity found in the digestive gland assays. The transcriptomic and biochemical data are consistent in suggesting that the cathepsin L-like cysteine proteinase encoded by C2952;N232;S21d is the primary proteolytic enzyme in the digestive processes of *D. reticulatum*. A previous EST study in *Diabrotica virgifera virgifera* showed high frequency of cysteine proteases in the gut tissue cDNA (Siegfried, Waterfield, and Ffrench-Constant 2005). These proteases have been implicated as key enzymes in plant-herbivore relationships, with cysteine proteases being a target for plant cystatins (Shindo and Van der Hoorn 2008). The presence of a second similar but distinct cathepsin L in Table 17, as well as several more partial homologues based on BLAST data may suggest a similar function to insects in *D. reticulatum*. Despite potentially a reliance on one key protease, a pool of alternative similarly functioning enzymes which could be utilised if the primary enzyme in use is inhibited.

4.2.2 Ferritin

An average of 1139 reads linked to the ferritin IPR terms in the digestive gland transcriptome analysis, in addition to a ferritin sequence being found in the random sequencing of cDNA clones conducted earlier (Ch. 3). Transcripts encoding

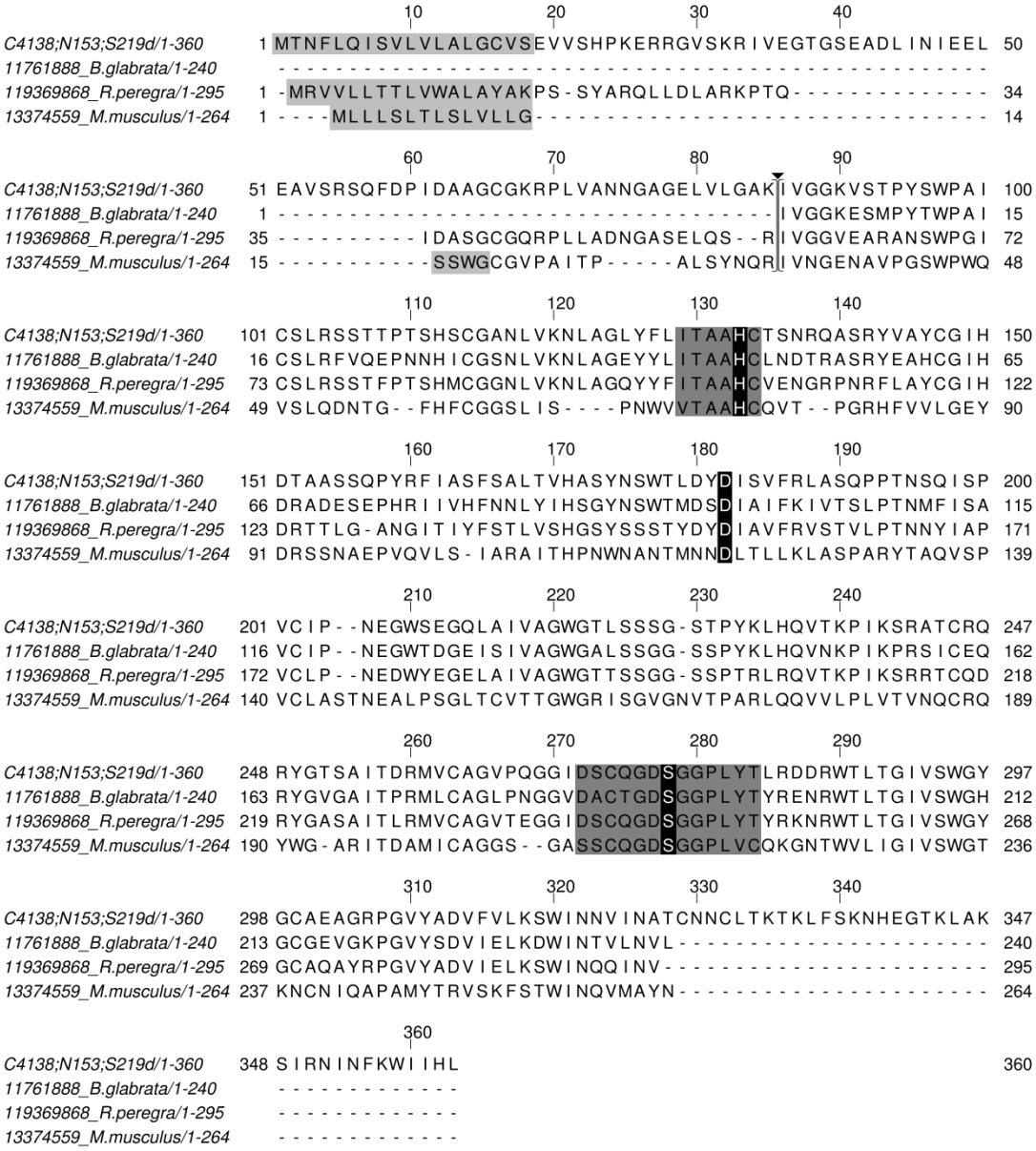


Figure 25 : Alignment of contig with serine protease homologues

Shows alignment of C4138;N153;S219d to 2 Molluscan and 1 chordate serine protease homologues from the NCBI nr database. Residues highlighted in black at His133 and Ser275 are conserved residues part of the PS00134 & PS00135 (highlighted dark grey), with PDOC00124 indicating a 100% likelihood that a protein containing these to patterns is a Serine Protease. Also highlighted black is a conserved Asp (~182aa) highlighted by the MEROPS database record for Chymotrypsin. Interpro returned the N-terminal SignalP prediction for the D. reticulatum contig as well as some homologues (highlighted light grey at the N-terminus) the bar and arrow represent the expected cleavage point of the propeptide region.

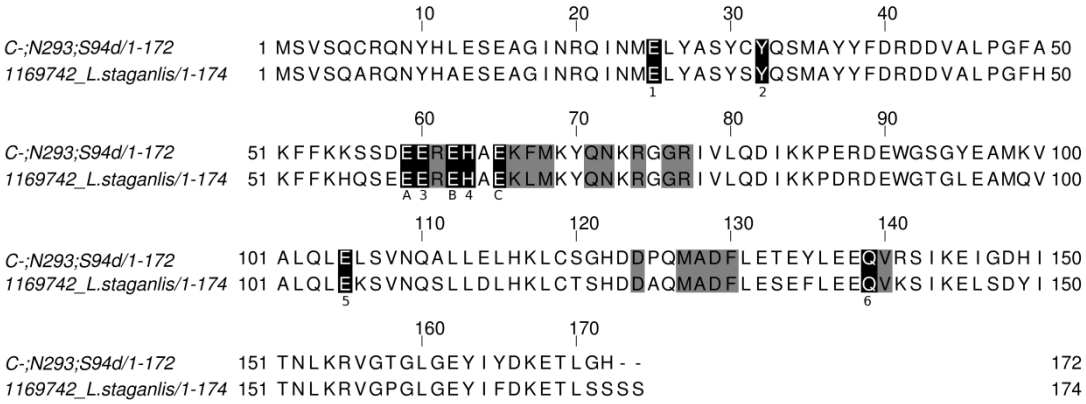


Figure 26 : D. reticulatum contig homologous to L. stagnalis snail soma ferritin

Alignment of C-;N293;S274d against L. stagnalis (CLCbio assembly does contain this high read count ferritin but is fragmented into 2 contigs). Highlighted amino acids A-C show 3 nucleotides of the iron nucleation site. Residues 1-6 are conserved residues of the H-specific ferroxidase centre. Additional highlighted in grey are 2 Prosite ferritin iron-binding region signatures attach to the ferritin conserved site IPR term (IPR014034). PS00540, E-x-[KR]-E-x(2)-E- [KR]-[LF]-[LIVMA]-x(2)-Q-N-x-R-x-G-R which includes the nucleation site, starting at residue A (Glu61(59)). PS00204 D-x(2)-[LIVMF]-[STACQV]-[DH]-[FYMI]-[LIV]-[EN]- x(2)-[FYC]-L-x(6)-[LIVMQ]- [KNER] includes Gln141(139) of the ferroxidase core and starts at Asp124

several different ferritin-like proteins are present in the transcriptome. The most significant contig in terms of reads encodes a protein most similar to Molluscan ferritins, having greatest similarity to *Lymnaea stagnalis* soma ferritin. In Mollusca 2 main types of ferritin have been identified; soma ferritin expressed constitutively in most tissues and yolk ferritin expressed primarily in the platelets of the growing oocytes (Darl et al., 1994). The *D. reticulatum* contig matching soma ferritin has ~1000 reads (C-;N293;S94d), compared with the most abundant sequence similar to yolk ferritin (C2570;N476;S435d), at 33 reads. Expression of soma and yolk ferritin has been shown to differ between tissue types, including the Molluscan midgut,

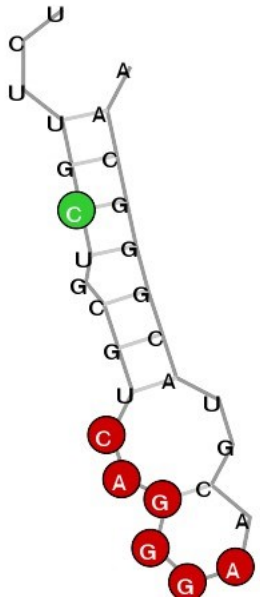


Figure 27 : SIREs RNA fold prediction for D. reticulatum Ferritin sequence
The green circle shows a potential C8 bulge usually found in IREs. The C here is neither at position 8 nor a bulge, but there is a G8 bulge present. The red circles show the predicted apical loop. [Image generated by SIREs webserver: http://ccbg.imppc.org/sires/index.html]

possibly due to differing regulation mechanisms (Darl et al., 1994; Bottke et al., 1988). The soma ferritin in *D. reticulatum* digestive gland is a highly abundant transcript, indicating a need for high levels of expression of the ferritin gene.

The ferritin complex, normally linked with iron storage and homeostasis (Crichton and Charloteaux-Wauters 1987), is typically assembled from 24 peptide subunits each containing an iron-binding domain. Conserved amino acids in ferritin include Asp131 and Glu134 within the supposed iron access channel formed by 3 neighbouring subunits, Glu61, 64, 67 forming the iron nucleation site and Glu27, Tyr34, Glu61, Glu62, His65, Glu107 and Gln141, conserved as part of the H-specific ferroxidase centre. These are all present in the *D. reticulatum's* soma ferritin sequence, see Figure 26. Like the 5' UTR of *L. stagnalis* mRNA encoding soma

ferritin, and the 5' UTRs of mRNAs encoding H-chain ferritin subunits generally, an Iron response element (IRE) is predicted to be present in the 5' UTR of *D. reticulatum* soma ferritin mRNA (C-;N293;S94d) based on SIREs analysis, an online tool for predicting the signature RNA folding of IREs (Campillos et al. 2010), see Figure 27.

Ferritin has been one of the most frequently sequenced transcripts in *D. reticulatum*, both with Sanger and 454 technologies. Despite this, the reason why it should be so common is not entirely clear as Molluscan ferritin has been linked with a number of functions. In the pearl oyster *Pinctada fucata* ferritin has been postulated to be important in shell formation. Expression data available shows similar levels of ferritin mRNA in the mantle and digestive gland of the organism, perhaps suggesting a translocation of iron, from the gut to the mantle via ferritin (Zhang et al. 2003).

However, the need for high levels of large iron translocation to the relatively small internal shell of *D. reticulatum* seems less likely. In the limpet *Cellana toreuma* ferric oxide has been shown to make up 54% of the radula, a series of rows of teeth used to graze on rock encrusted algae (Lu, Huang, and Li 1995). New rows of radula teeth are formed through biomineralisation with ferritin present in the tooth surrounding cells and the teeth themselves. Subcellular investigation demonstrated ferritin was transported and disassembled as the tooth matured with final iron deposition on the tooth cusp (Clark et al. 1995). The radula in *D. reticulatum* may show similar mineralisation, and will suffer high rates of wear from abrading plant tissue; a high turnover of teeth could explain the high levels of ferritin mRNA in the digestive gland transcriptome, with ferritin used for scavenging and storing iron from the diet.

4.3.3 Glycosidase Hydrolase Family 9

Glycosidase Hydrolase Family 9 (GHF9) is a family of enzymes which includes most cellulases, and has previously been known as cellulase family E. GHF9 cellulases have been identified previously in Mollusca, and can be loosely grouped into the cellulases (EC 3.2.1.4) or cellulose 1,4-β-cellobiosidase (EC

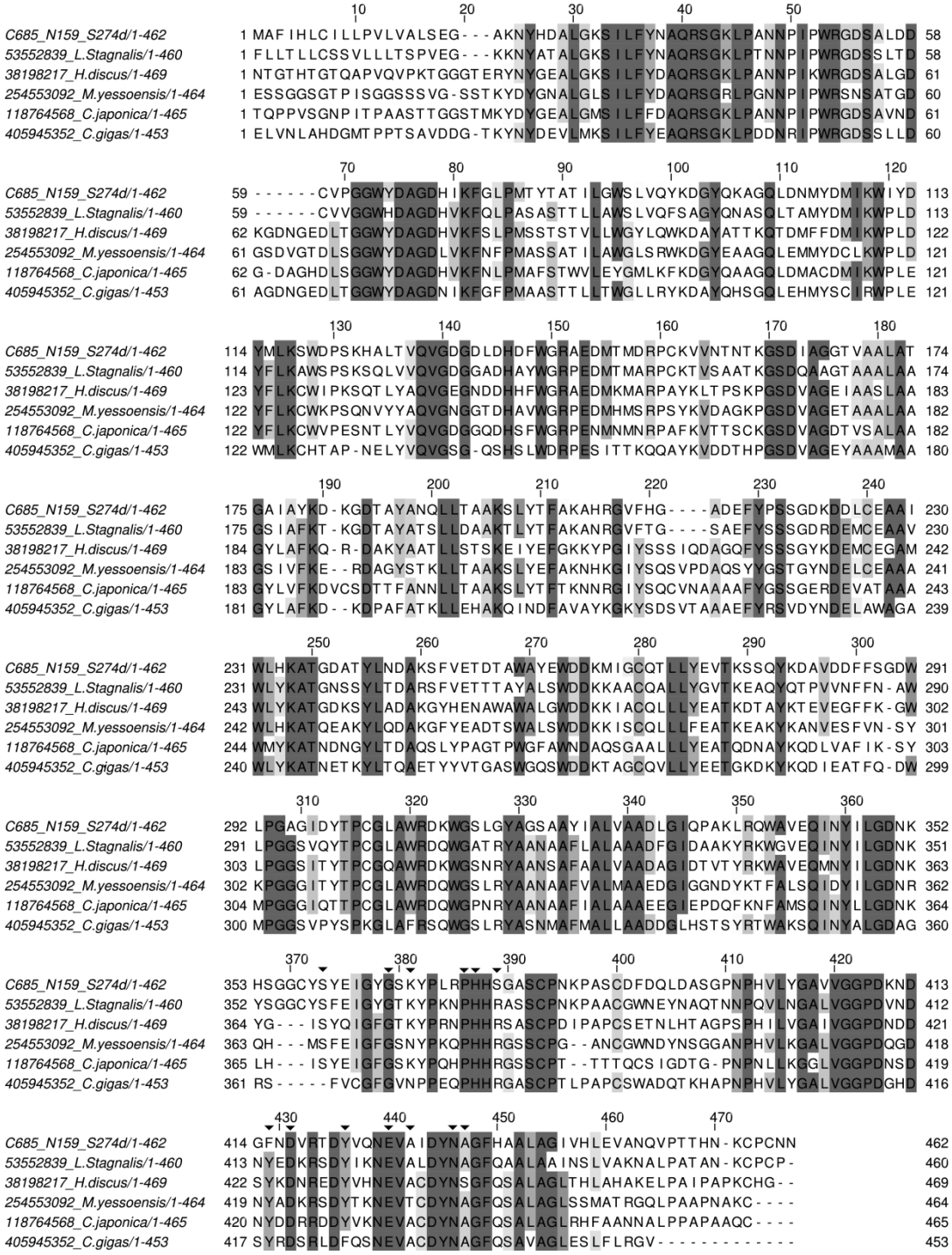


Figure 28 : Alignment of D. reticulatum contig with Molluscan homologues

Figure shows the GHF9 alignment generated by comparing D. reticulatum with other Molluscan sequences. The two areas with black triangles highlight the most conserved regions linked to cellulases by IPR terms. The first pattern (PS00592) covering the consensus region 374-390 contains the active site histidine. The second highlighted pattern (PS00698) covers the consensus

region 430-448 and contains two important catalytic residues: an aspartate and glutamate. The 5' region of M. yessoensis has been truncated for brevity with the sequence beginning at residue 138 which is shown here as residue 1. Several other Molluscan sequences not shown here have extended 5' regions as with M. yessoensis, although comparison show the region to be variable with poor homology between the Molluscan species.

3.2.1.91) classification. In both cases these enzymes hydrolyse (1→4)-β-D-glucosidic linkages in cellulose and cellulose-like polysaccharides such as cereal β-D-glucans; however, specificities differ in that cellulose 1,4-β-cellobiosidase hydrolyses the non-reducing end of the polysaccharide releasing cellubiose, whilst cellulase is an endohydrolase, cleaving glycosidic bonds within the polysaccharide chain. *D. reticulatum* has been shown to have cellulase activity by biochemical assay (Stone and Morton 1958), and therefore IPR terms IPR001701 and IPR018221 would be expected to be present in this family of glycosidases found within the digestive gland transcriptome.

An average of 1314 reads are linked to the GHF9 active site. This IPR term group has high prevalence within the assembly, being the second highest ranked IPR term, below cysteine proteases. C685;N159;S274d is the largest sequence linked to this IPR term, appearing to be a full length EST containing an ORF corresponding to a 462 residue peptide. This peptide shows similarity to a number of Molluscan cellulases including Abalone (*Haliotis discus)*, Scallop (*Mizuhopecten yessoensis)*, Clam (*Corbicula Japonica)* and Snail (*Ampullaria crossean)*. 2 Prosite patterns are linked to this IPR term, PS00592 [STV]-x-[LIVMFY]-[STV]-x(2)-G-x-[NKR]-x(4)-[PLIVM]-H-x-R with H being an active site residue and PS00698 [FYW]-x-D-x(4)-[FYW]-x(3)-E-x-[STA]-x(3)-N-[STA] with D & E as active site residues. Both patterns are observable in the *D. reticulatum* sequence, but with 2 substitutions near the beginning of PS00592, see Figure 28. Substitutions at these points are conserved within but not outside of the Molluscan phylum, including *H. discus*, though it has a Gln in place of Glu.

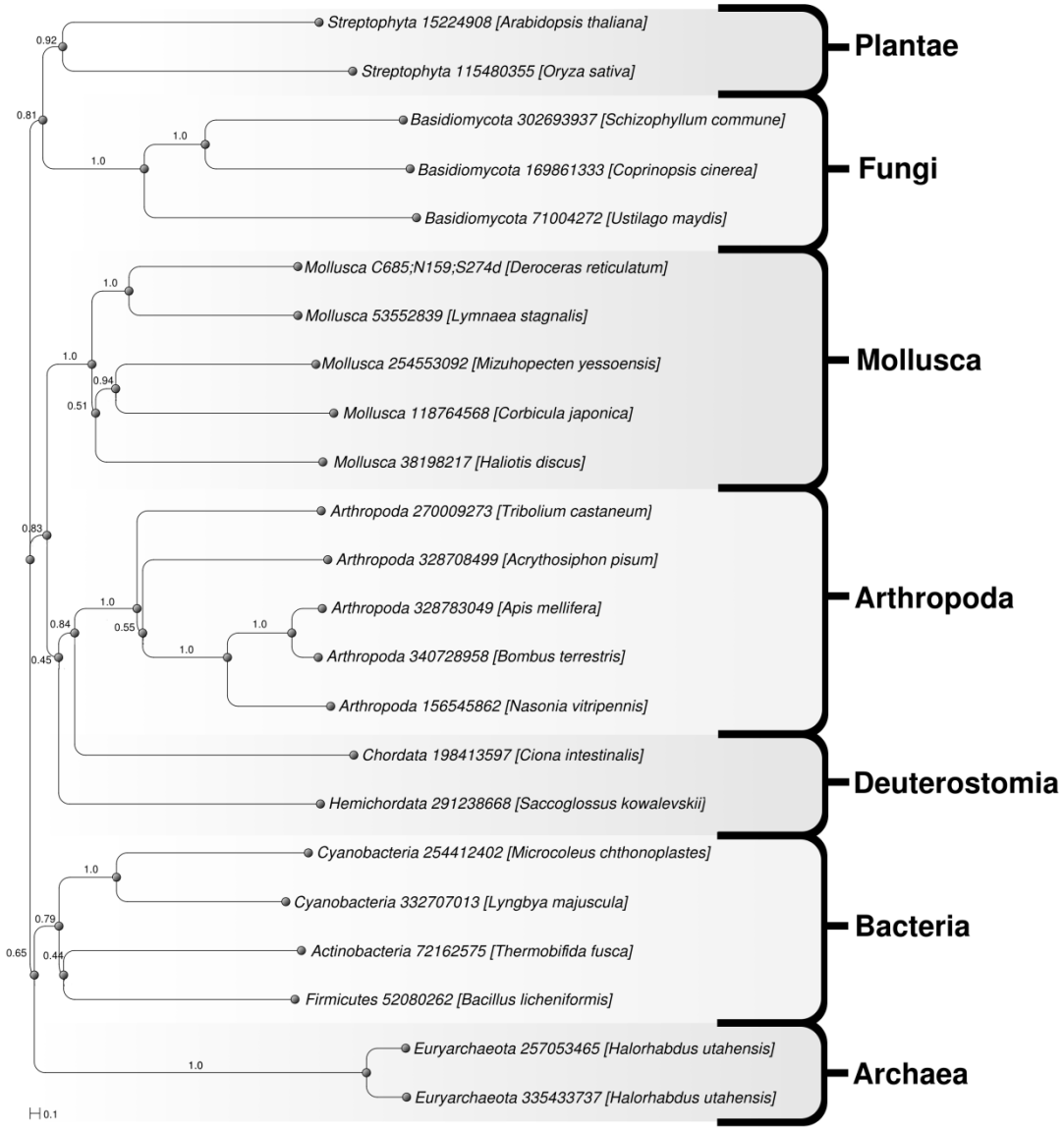


Figure 29 : Phylogenetic tree of D. reticulatum GHF9 with homologues

The tree showing the closest GHF9 homologues from various species (labelled as Phylum GI No [Species]), with additional brackets to highlight branches. The species groups have been highlighted and are positioned as would be expected.

Figure 29 shows the relationship between the sequence encoded by C685;N159;S274d, a full length GHF9 enzyme, and most similar sequences from other phyla, including Mollusca. The Molluscan sequences group together with high

bootstrap confidence, with *L. stagnalis* being closest homologue to *D. reticulatum*. This phylogenetic tree has a strict grouping of sequences with phyla, with no species breaking the pattern. The current opinion is that all cellulases were lost from many metazoan species, rather than known metazoan cellulases being the product of horizontal gene flow (Davison and Blaxter 2005). In highlighting the position of *D. reticulatum* cellulase on the phylogenetic tree by sequence comparison, these results suggest the same conclusion.

4.4 Other notable groups

4.4.1 C-Type Lectins

C-Type lectins are a diverse group of proteins with varying functions which share a conserved domain; this normally has carbohydrate binding affinity, but can have other functional roles. Sequence motifs are associated with carbohydrate binding activity. The C-Type lectin (CTL) IPR term (IPR001304) is linked to 14 contigs, 7 of which appear to be full coding sequences for single CTL domain containing proteins. All but 1 of the predicted sequences are linked to the PROSITE pattern PS00615 which describes the carbohydrate binding domain site; the unlinked peptide still contains the canonical cysteine positions of the C-Type Lectin domain, and a WND-like motif associated with carbohydrate binding (see Figure 30). The predicted sequences can be divided into 2 types; those containing a motif of 3 conserved residues, and those without. This motif, which is also referred to as the QPD/EPN motif, is a major determinant of sugar binding specificity, since conversion of the EPN sequence to QPD increases affinity for galactose binding (Drickamer 1992).

C-type Lectins are classified into 15 families or groups based on domain structure. Only 6 groups are proteins containing a single CTL domain and no other functional domains (Zelensky and Gready 2005); of the 6 single domain CTL groups, 4 contain transmembrane regions. Despite several *D. reticulatum* contigs linking to PANTHER terms (via INTERPRO) (PTHR22801:SF15) for groups containing transmembrane regions (primarily Group II, Asialoglycoprotein and DC receptors) none of the *D. reticulatum* C-type lectin contigs are predicted to contain these

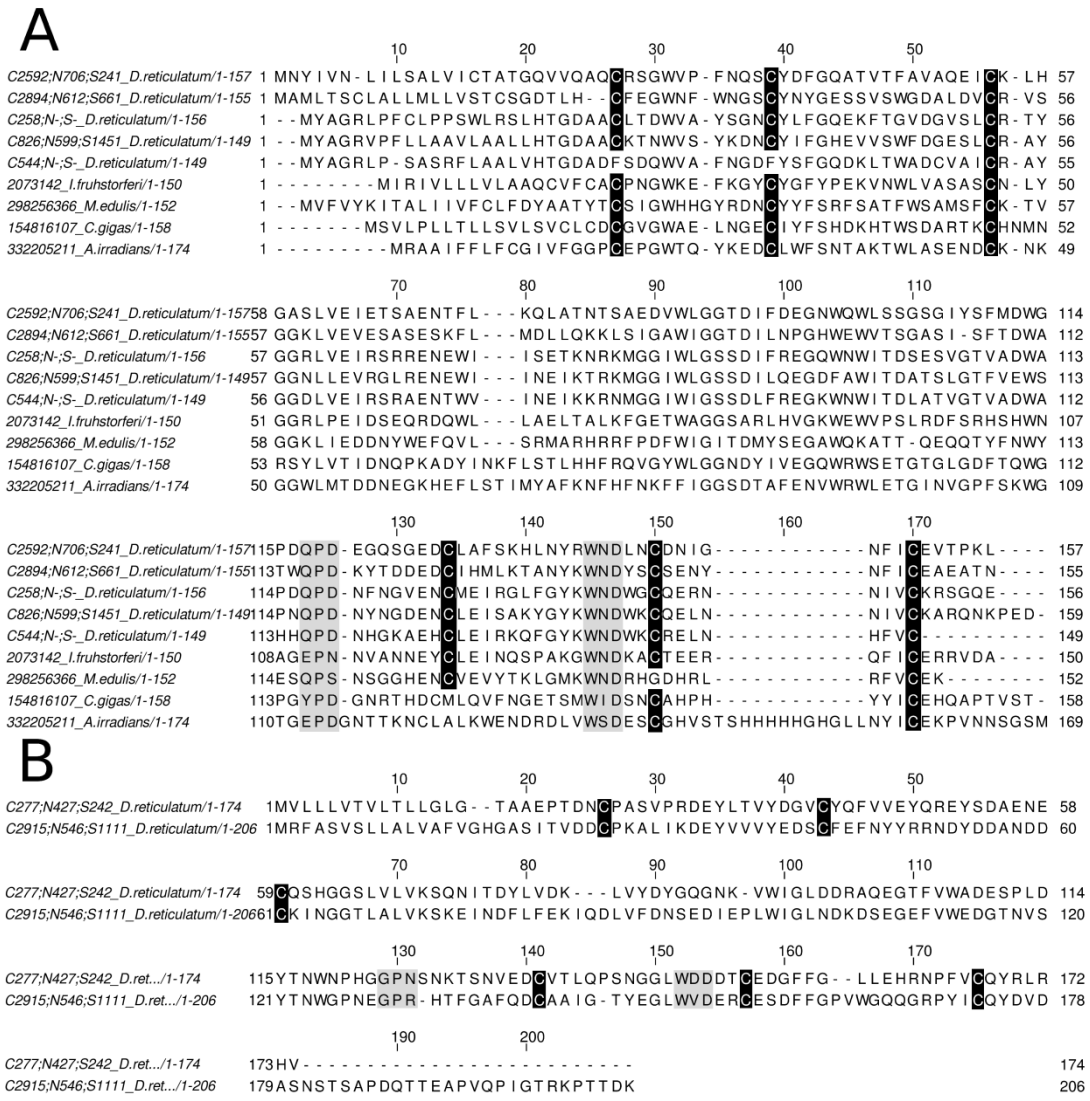


Figure 30 : Sequence alignment of C-Type lectins

A) The alignment of 5 QPD motif containing contigs which match the C-Type lectin IPR001304 term, and are shown alongside several Molluscan C-Type lectins for comparison. Cysteine residues form disulphide bridges at alignment positions 56-170 and 134-150, usually termed the long and short bridges. Also shown is the N-terminal 27-39 disulphide bridge, which is notably missing from C5444;N-;S-d. The WND motif is highly conserved between C-Type lectins and, the QPD motif suggests a higher affinity for galactose. 2 of the D. reticulatum sequences have no SignalP signalling peptide predictions.

B) 2 further contigs which do not contain the QPD motif, but have conserved cysteines and WND-like domains and SignalP prediction. Contig3176 contains no observable acidic residue in the region of the QPD/EPN motif usually seen, an indicator in mammalian C-Type lectins of loss of Carbohydrate binding.

features, although it is possible that some of the predicted *D. reticulatum* CTLs could be partial sequences of larger proteins. The 2 CTL groups with a single domain and no transmembrane regions are REG and eosinophil major basic proteins; neither have significant homology to *D. reticulatum* contigs, nor do their functions seem likely to be similar in Mollusca. The classification systems for CTLs have mainly been defined through mammalian genomes, and may not be appropriate for Molluscan sequences. The single domain CTLs identified in *D. reticulatum* are most similar to other Molluscan predicted proteins.

In Pacific Oyster (*C. gigas*), 2 CTLs were isolated, 1 almost solely expressed in the digestive gland. The function suggested was a role in the immune defence of the organism, with the secreted CTL in digestive gland potentially used for pathogen recognition, though no up-regulation was seen during biological challenge (Yamaura, Takahashi, and Suzuki 2008). In Blue Mussel (*Mytilus edulis*), a CTL was isolated whose expression was tied to the starvation/feeding of the organism, leading to a suggested function of food particle recognition, important for a filter feeding animal (Pales Espinosa, Perrigault, and Allam 2010). CTLs have been isolated from mucous of a number of animals, with Molluscan mucus also containing single domain CTLs, which were shown to have agglutination ability as well as showing similarity to antifreeze protein (Yuasa et al. 1998).

There are significant similarities between many of these proteins and the *D. reticulatum* CTLs, and 4 of the *D. reticulatum* CTLs containing QPD motifs show significant similarity, using the InterProScan tool, to anti-freeze proteins (IPR002353). However, in a phylogenetic tree of Molluscan CTLs (Figure 31) it is immediately obvious that despite sharing a common CTL domain, evidence that any of these proteins are functional homologues is limited, due to early branching with low bootstrap values. The functional roles assigned to many of these protein sequences have been derived from sequence homology, without any clear linkage to a protein whose function has been determined. Attempting to describe a function for Molluscan CTLs based on sequence similarity is unlike similar analyses for other proteins described in this paper, where a clear indication of the likely functional role could be derived from sequence similarity to an enzyme or protein whose function

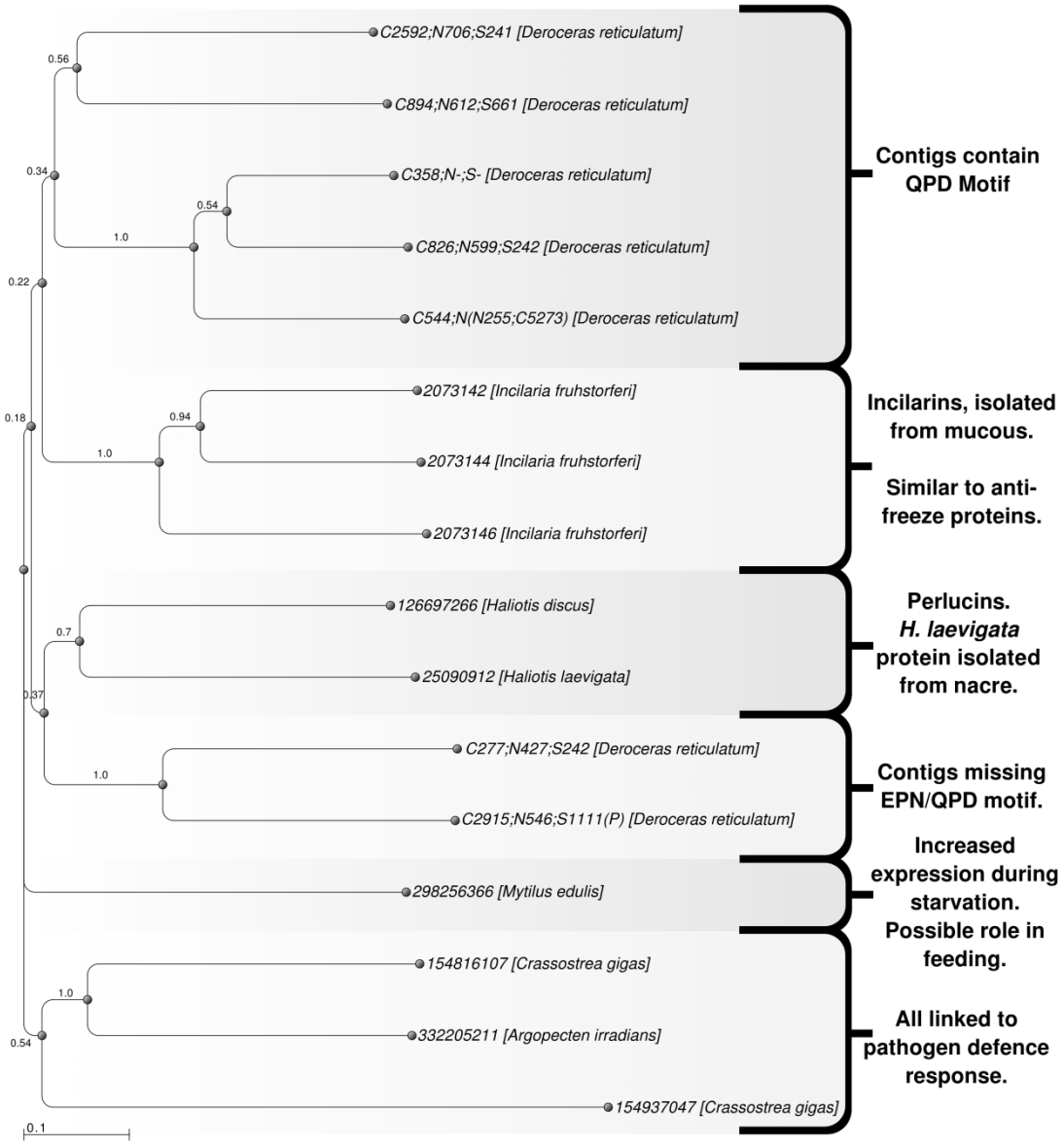


Figure 31: Phylogenetic tree of Deroceras reticulatum and C-Type Lectin homologues

A phylogenetic tree produced from the alignment of 7 contigs linked to the Interpro term for C-Type lectin (CTL) (IPR001304) as compared with a number of Molluscan CTLs, labelled as GI no. and species. Sequences which have similar functions group together. However sequences from D. reticulatum do not group particularly well with many contigs having low bootstrap and early branching. In general we see these groups of CTLs as relatively divergent and as such do not suggest any function for D. reticulatum sequences.

was well characterised, and in addition biochemical evidence was available to support the presence of the function. Suggestion of functional roles for the *D. reticulatum* CTLs will require independent evidence, such as biochemical analysis or functional characterisation, to validate the predictions.

4.4.2 P60-like

Whilst limited information can be discerned about this gene it is present in all assemblies, though highly significant in CLCbio and shows strong enough homology to other sequences from other species to indicate it's an actual gene. P60-like (IPR011687) is the linked to a single contig which is a partial sequence, similar to Glioma Tumour Suppressor Candidate Region Gene 2 (GLTRSCR2). Despite this gene being described as a tumour suppressor gene, the annotation is based on the gene locus being within a candidate region for a tumour suppression gene on human chromosome 19q (Smith et al. 2000) rather than functional characterisation, although evidence for its involvement in regulating apoptotic processes in human cells has been presented (Yim et al. 2007). The IPR term is linked with 2 database records, P60-Like from the Panther database and Nop53 from the PFAM database.

Nop53 is a protein found in most Eukarya from yeast to humans and is involved in late stage rRNA processing as well as having a role in ribosome biogenesis (Granato et al. 2005). Deletion of the gene in yeast to produce null mutants severely impairs their growth rate, limiting cell viability (Sydorskyy et al. 2005). The level of similarity between the *D. reticulatum* sequence and Nop53 proteins from other organisms is limited compared to some well conserved proteins, but several regions show higher conservation. Considering regions of similarity to Nop53, the most likely function of the coding sequence is related to rRNA processing and biogenesis, although the reason for the abundance of this transcript in one assembly is unclear.

4.4.3 Other Glycoside Hydrolases

There is a wide range of glycoside hydrolase families (GHF). These enzymes can be defined as catalysing reactions hydrolysing *O*- and *S*-glycosidic bonds in a

IPR Term	GF	Enzymes in Family	Average Reads
IPR001701	9	endoglucanase (EC 3.2.1.4); cellobiohydrolase (EC 3.2.1.91); β-glucosidase (EC 3.2.1.21); exo-β-glucosaminidase (EC 3.2.1.165)	1374
IPR026892	3	β-glucosidase (EC 3.2.1.21); xylan 1,4-β-xylosidase (EC 3.2.1.37); β-N-acetylhexosaminidase (EC 3.2.1.52); glucan 1,3-β-glucosidase (EC 3.2.1.58); glucan 1,4-β-glucosidase (EC 3.2.1.74); exo-1,3-1,4-glucanase (EC 3.2.1.-); α-L-arabinofuranosidase (EC 3.2.1.55)	86
IPR000322	31	α-glucosidase (EC 3.2.1.20); α-1,3-glucosidase (EC 3.2.1.84); sucrase-isomaltase (EC 3.2.1.48) (EC 3.2.1.10); α-xylosidase (EC 3.2.1.177); trehalose-6-phosphate hydrolase (EC 3.2.1.93); oligo-a-glucosidase (EC 3.2.1.10) ...	63
IPR015902	13	α-amylase (EC 3.2.1.1); pullulanase (EC 3.2.1.41); cyclomaltodextrin glucanotransferase (EC 2.4.1.19); cyclomaltodextrinase (EC 3.2.1.54) ...	60
IPR000933	29	α-L-fucosidase (EC 3.2.1.51); α-1,3/1,4-L-fucosidase (EC 3.2.1.111)	35
IPR001223	18	chitinase (EC 3.2.1.14); endo-β-N-acetylglucosaminidase (EC 3.2.1.96); xylanase inhibitor; concanavalin B; narbonin	32
IPR000757	16	xyloglucan:xyloglucosyltransferase (EC 2.4.1.207); keratan-sulfate endo-1,4-β-galactosidase (EC 3.2.1.103); endo-1,3-β-glucanase (EC 3.2.1.39)...	18

IPR015883	20	β-hexosaminidase (EC 3.2.1.52); lacto-N-biosidase (EC 3.2.1.140); β-1,6-N-acetylglucosaminidase (EC 3.2.1.-); β-6-SO3-N-acetylglucosaminidase (EC 3.2.1.-)	3
IPR001944	35	β-galactosidase (EC 3.2.1.23); exo-β-glucosaminidase (EC 3.2.1.165)	1

Table 18: Table of all glycosidase family IPR terms represented in the digestive tissue data

Table includes the IPR terms and their related Glycosidase Family number, all of which fit into the CAZy annotation of EC 3.2.1.. Enzyme names and CAZy annotation, appended in parenthesis, demonstrate enzymes found in each group. Families 31, 13 and 16 have been abridged for brevity but all enzymes can be found at the CAZy website.*

range of compounds, and have the EC number 3.2.1.*. So far 85 different families have been defined based on the sequence similarity, though these families can themselves contain broad subgroups; an extensive online database describes these families and their known members (http://www.cazy.org/Glycoside-Hydrolases.html). A number of useful predictions can be made as catalytic machinery and molecular mechanism is conserved for the majority of the GHFs (Gebler et al. 1992). 9 IPR terms linked to the GHFs can be found in the *D. reticulatum* digestive gland transcriptome, as shown in Table 18. Many of the corresponding enzyme activities were previously shown to be present in the *D. reticulatum* digestive crop juice (Runham and Hunter 1970). Homologues to GHF9, cellulase, have been sequenced in over 10-fold in abundance relative to other GHF enzymes. This may indicate this cellulase enzyme as having a more significant role, with other hydrolases being less important or alternatively are many times more efficient than the cellulase homologue.

In *D. reticulatum* a single incomplete contig is homologous to GHF3, the top homology match to 'Putative β-D-xylosidase 2' in *C. gigas*. The GHF3 family

currently groups together as exo-acting β-D-glucosidases, α-L-arabinofuranosidases, β-D-xylopyranosidases and *N*-acetyl-β-D-glucosaminidases (Harvey et al. 2000) many of which are referred to as hemicellulases. These enzymes are widely distributed in bacteria, fungi and plants with enzymatic functions involved in cellulosic biomass degradation, plant and bacterial cell wall remodeling, energy metabolism and pathogen defense. The enzymes in this family are of particular interest in biotechnology as tools for the production of bioethanol. Enzymes including β-glucosidases and xylosidase enzymes are involved in the degradation of xylans by wood-degrading organisms (Fujii et al. 2009). Many of these GH3 enzymes act as accessory enzymes in xylan degradation, specifically digesting particular residues, such as such as arabinofuranosidases which hydrolyse L-arabinose linkages found in many polysaccharides (Numan and Bhosle 2006).

However in metazoa, GHF3 enzymes are poorly characterised, though predicted homologues from this group can be found in many animal phyla. Previous studies in *D. reticulatum* and other Molluscan species showed xylanase activity (Nielsen 1962; Stone and Morton 1958). But in many other invertebrates studied, such as termites and grasshoppers, xylanase activity is performed by gut microbiota (Shi et al. 2013; Bastien et al. 2013). It seems highly probable that *D. reticulatum* would have endogenous enzymes to hydrolyse β-glucans. But as far as our research indicates no studies have yet to resolve functionality of invertebrate homologues, beyond the homology to the GHF3 group as a whole. As such predictions are limited to enzyme functionality that acts on β-glucans, and without the N-terminal region of the contig this could be either a digestive enzyme or alternatively a cytosolic enzyme as found in humans (Dekker et al. 2011).

GHF31, like GHF3, also includes glucosidase enzymes, but this group hydrolyses O-glycosidic bonds in α-glucan, which include dextran, glycogen, pullulan, and starch. Due to a greater level of characterisation of metazoan enzymes of this group a greater detail of prediction can be made. Additionally the contig, although missing the C-terminus has a predicted signal peptide which would indicates it as having a potential extracellular enzymatic function. The most likely function is that of α-glucosidase (EC 3.2.1.20) due to strongest homology to other α-

glucosidase 2 proteins based on BLAST nr and panther database sequence comparisons. A lack of a transmembrane region indicates the sequence is less likely to be related to sucrase-isomaltase (EC 3.2.1.48) (EC 3.2.1.10).

GHF13 is a wide ranging of enzymes acting on polysaccharide substrates containing α-glucoside linkages. Whilst α-amylases are the best described in the group, it is one of the most diverse groups containing 22 EC categories, with studies subdividing it further to 35 subfamilies (Stam et al. 2006). The main contig, C2633;N151;S537d, in this group appears to be a complete coding region with a predicted signal peptide for which the top BLAST nr match is Maltase A6 in *D. melanogaster* (GI:221330053), see Figure 32. Maltase A6 flybase record (FBgn0050360) shows very high expression in the adult fly midgut and only low level expression in none digestive-related tissues. Both the contig and the *D. melanogaster* gene are annotated by BLAST CDD as a α-phosphotrehalase (trehalose-6-phosphate hydrolase, EC 3.2.1.93). However key conserved amino acids from previous enzymes that act on trehalose-6-phosphate are not present in *D. reticulatum*. Amino acids 207-210, relative to Taka Amylase A (GI: 23586) included for positional reference (Matsuura et al. 1984), in the *D. reticulatum* are Ala-Ile-Gln-Gln. These residues are not seen in enzymes with trehalose substrate, but are the equivalent to residues in α-amylases which act only on α-1,4-bonds (MacGregor, Janeček, and Svensson 2001).

Both GHF 29 and 35 contain a much smaller range of enzymatic functions. GHF 29 contains exo-acting α-fucosidase with no other activities described. Notably this gene has been functionally characterised in a mollusc *Pecten maximus*, where hydrolysis of fucosyl units of fucoidan was demonstrated (Berteau et al. 2002). The same group had previously shown numerous other glycosidase functions in this species (Daniel et al. 1999), but to date none of these functions have been linked to transcript or peptide sequence data. The majority of GHF 35 enzymes are β-galactosidases which act on β-1,3-, β-1,6- or β-1,4-galactosidic linkages, with the remaining enzymes exo-β-glucosaminidases only been found in archaea (Liu et al. 2006), though exo-β-glucosaminidases from other GHFs have been found in bacteria (Nanjo, Katsumi, and Sakai 1990) and fungi (Ike et al. 2006).

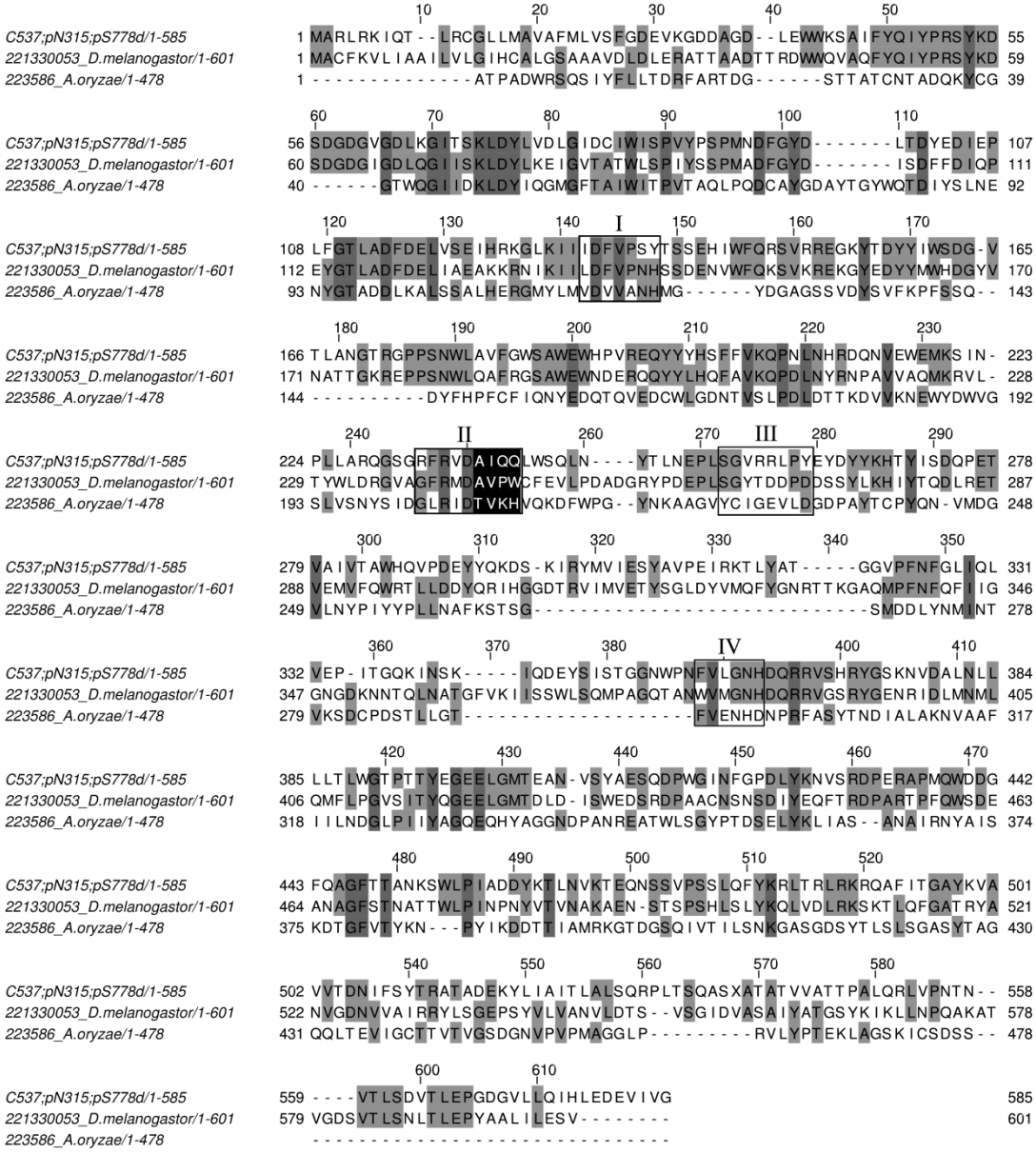


Figure 32 : Alignment of D. reticulatum contig with GHF13 homologues

Alignment of C2633;N151;S537d with top matching homologue from NCBI nr database Maltase A6 from D. melanogaster. Also included is Taka Amylase A from Aspergillus oryzae as a reference for key residue positions. Boxed areas represent 4 conserved regions which should maintain similar degrees of hydrophobicity with homologues. 4 residues (white letters on black) represent key positions which correlate to substrate binding specificities in GHF13 enzymes. In this the D. reticulatum, Ala-Ile-Gln-Gln residues appear more similar to α-amylases. D. melanogaster's Ala-Val-Pro-Trp is most similar to Ala-Val-Pro-Tyr associated with trehalose synthase.

GHF18 enzymes include chitinases (EC 3.2.1.14) and endo-β-N-acetylglucosaminidases (EC 3.2.1.96) as well as proteins with no hydrolysis function such as lectins and xylanase inhibitors. Chitinase has previously been shown to be present in the crop juice of *D. reticulatum;* as such it seems probable that some of the GHF18 sequences maybe digestive chitinase. Interestingly unlike other GHFs, there is a similar number of reads attached to the GHF18 in the neural dataset, average of 21 for IPR001223, suggesting that some enzymes may not be specific to digestion. Several contigs match IPR001223 and IPR001579, Chitinase II and Glycoside hydrolase chitinase active site, and have strongest BLAST homology to other chitinase genes. One of the 2 chitinase contigs in the digestive gland is also present in the neural tissue dataset.

Chitinase has previously been shown to be present in the crop juice of *D. reticulatum;* as such it seems probable that some of the GHF18 sequences maybe digestive chitinase. However in the neural tissue whilst matches to GHF9 active site are absent, glycoside hydrolase chitinase active site links in the neural dataset are few but notably not absent. With chitinase being identified in the saliva of octopus *Eledone cirrhosa* (Grisley and Boyle 1990); this could potentially be contamination from the salivary gland, which is spatially close to the nervous system. Alternatively previous studies in molluscs have shown the both a ubiquitous low expression pattern of chitinase in *C. gigas*, and chitinase and chitinase-like enzymes posited to be related to early development and immune processes (Badariotti, Thuau, et al. 2007; Badariotti, Lelong, et al. 2007). These studies indicate molluscs have additional roles for chitinases outside of the digestive gland. However this protein contained a chitin binding domain and predicted trans-membrane region. With none of the *D. reticulatum* GHF18 contigs being complete, comparison with the *C. gigas* enzyme cannot reliably be made. Despite this, there is a great deal of interest in chitinase and its potential use in biotechnology. Chitin is the second most abundant carbohydrate after cellulose and there is increasing number of applications for enzymes which degrade these polysaccharides into simpler sugars (Yuting Zhang et al. 2013); as such these enzymes may be of interest for future research.

4.5 BLAST homology analysis of Neural tissue data

Sequences from the transcriptomic analysis of neuronal tissue from *D. reticulatum* were assembled and analysed as described for sequences from digestive gland. Contigs from each assembly were compared with the NCBI nr database as was done for the digestive gland, Table 19 summarises the results and Figure 33 shows how the number of hits to the NCBI database changes depending on e-value cut-off. SeqMan and CLCbio have a very similar number of BLAST matches against contigs as well as reads, but give vastly different results in number of homologues identified. This may suggest the similarity is coincidental, with SeqMan being more fragmented whilst CLCbio giving better representation of different transcripts, possibly related to the difference in read inclusion seen in the assembly statistics, Table 12. Top phyla matches rank in the same order as digestive gland, with

Summary (All expect < 1e-3)	CLCbio	Newbler	SeqMan
Contigs with BLAST matches	1463	744	1586
Total NCBI protein homologues	44483	21678	30350
Number of reads with hits	29010	37249	28356

Table 19: Summary of BLAST homology matches for neural tissue assemblies

Mollusca representing between 27-38% of the top BLAST hits, followed by Chordata and Arthopoda, see Figure 34. The only difference in ranking between the two tissue types appears to be Echinodermata, but this is of limited significance as in both tissue types representation is around 2%, the value used as a cut-off for inclusion in the others category. Species are more variable between assemblers than seen in digestive gland tissues, with SeqMan NG having a notable increase in matches to other species, Figure 35. The types of species and their number are generally the same, although *A. californica* has a higher significance in neural tissue when

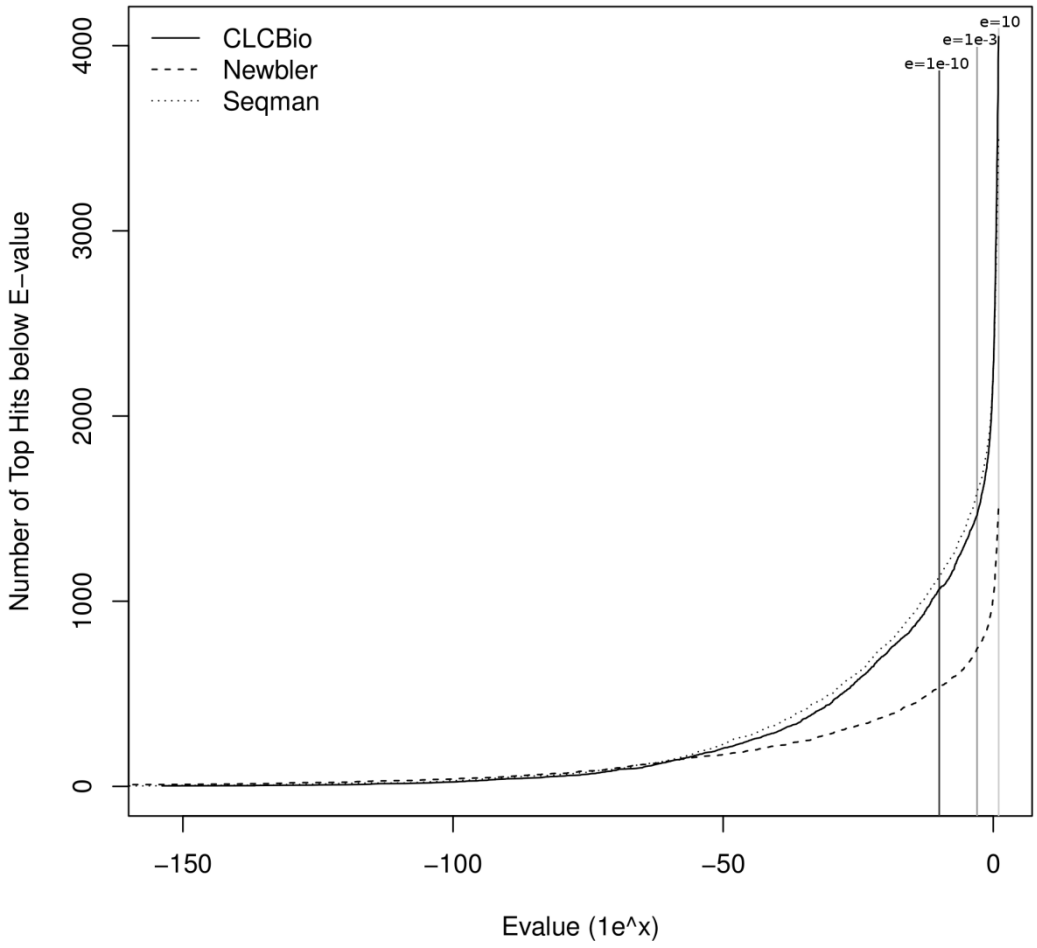


Figure 33 : Number of top BLAST hits below BLAST evalue for 3 assemblies

The graph represents the change in number of contig matches for neural tissue BLAST results which would be filtered out of the results depending on the e-value cut-off. Expect values of 1e-10, 1e-3 and 10 are shown as vertical lines with the former 2 representing common cut-offs and the later representing the limit of expect value for a BLAST search.

considered relative to other species, such as N. vectensis and B. floridae. This may reflect the use of A. californica as a model neurological system to a certain degree.

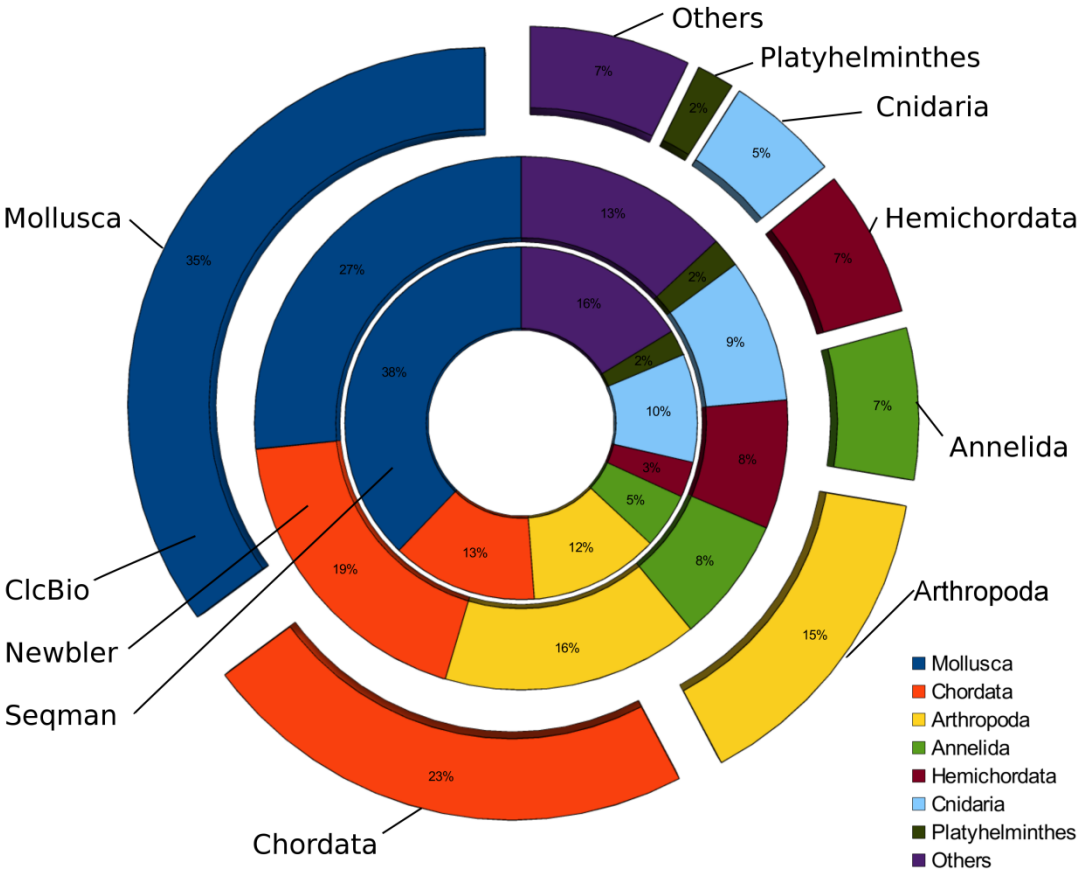


Figure 34 : Top BLAST matches per phyla as a percentage for the neural tissue

The overall pattern of phyla distribution is very similar to the digestive gland. With Mollusca with the most top hits as expected, followed by Chordata and Arthropoda.

4.5.1 INTERPRO Analysis of Neuronal Tissue

Table 20: Summary of top IPR terms for neural tissue assemblies summarises some of the IPR terms attached to greatest number of reads. Results from the different assemblies do not show as much similarity as was seen with digestive gland tissue, although all agree on the myosin signature as the top IPR term in for all assemblies. However, as described in the introduction, most of the neural tissue is associated with connective tissues. As described in the methods section, for neural

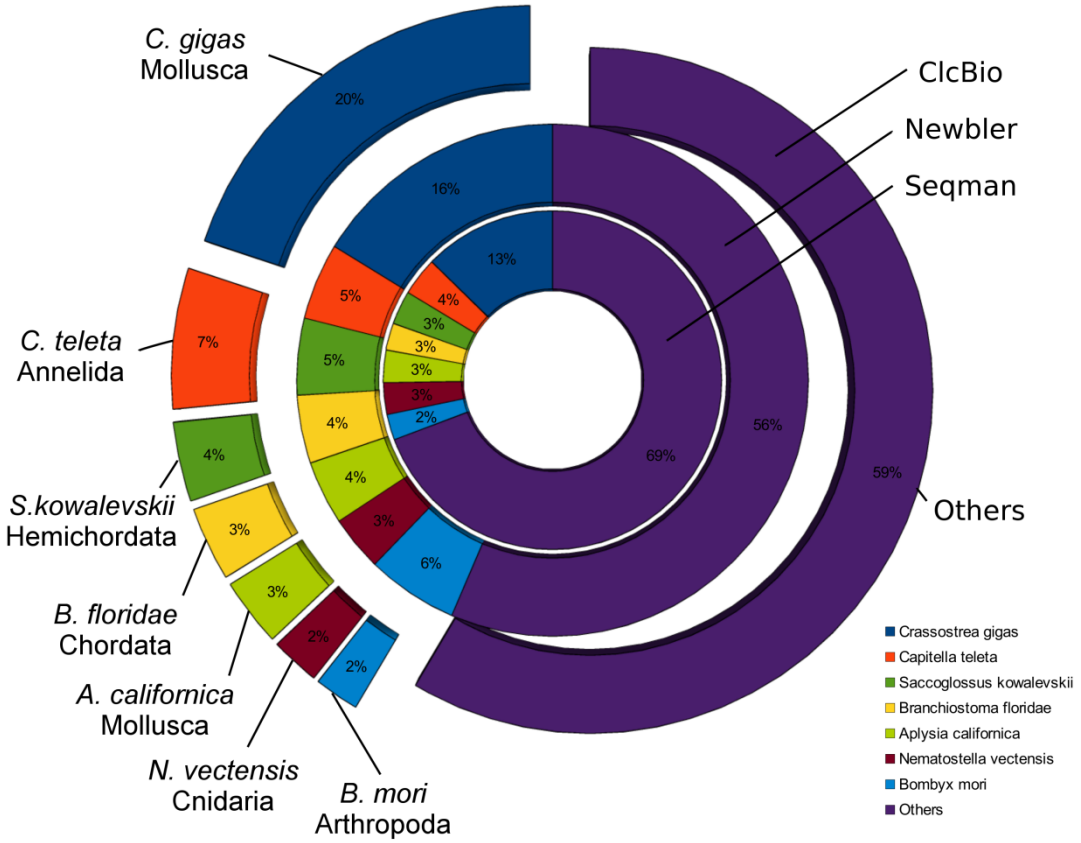


Figure 35 : Breakdown of top hits by species for the neural tissue assemblies

Other species take a greater proportion than with digestive gland species, though in general the same species are top. A. californica is a model organism for neurological research and here represents a slightly large portion than with digestive gland. Additionally B. mori represents quite a large share, particularly for the Newbler assembly, the reason for which is unclear.

tissue, to maximise tissue weight, sensory tentacles were also included within the RNA extraction due to the digitate ganglion and the optic and olfactory nerves connected from the sensory tentacles to the nerve ring. Thick retractor muscles are a requirement for variable orientation and fast retraction of sensory tentacles to avoid damage. As such the presence of myosin as well as connective tissue related proteins such as actin would be expected to be highly expressed transcripts.

CLCBIO

Main Term	Interpro Group	Read Count
IPR002928	Myosin tail	2562
IPR016021	MIF4G-like (Nuclear Cap Binding Protein)	2435
IPR004000	Actin	795
IPR020568	Ribosomal protein S5 domain	737
IPR013320	Concanavalin A-like lectin/glucanase	722
IPR001580	Calreticulin/calnexin	687
IPR001404	Heat shock protein Hsp90	626
IPR002130	Cyclophilin-like peptidyl-prolyl cis-trans isomerase domain	625

NEWBLER

Main Term	Interpro Group	Read Count
IPR002928	Myosin tail	2562
IPR016021	MIF4G-like (Nuclear Cap Binding Protein)	2435
IPR004000	Actin	795
IPR020568	Ribosomal protein S5 domain	737
IPR013320	Concanavalin A-like lectin/glucanase	722
IPR001580	Calreticulin/calnexin	687
IPR001404	Heat shock protein Hsp90	626
IPR002130	Cyclophilin-like peptidyl-prolyl cis-trans isomerase domain	625

SEQMAN

Main Term	Interpro Group	Read Count
IPR002928	Myosin tail	1799
IPR005819	Histone H5	1221
IPR020568	Ribosomal protein S5 domain	630
IPR001404	Heat shock protein Hsp90	587
IPR014764	Defective-in-cullin neddylation protein	556
IPR004000	Actin	485
IPR002048	Calcium-binding EF-hand	455
IPR002557	Chitin binding domain	382

Table 20: Summary of top IPR terms for neural tissue assemblies

In all cases myosin tail is the top IPR term for neural tissue. Hsp90 is also significant with similar read counts for all assemblies as is Ribosomal protein S5. However there are a number of IPR terms which only exist with high read counts in 1 or 2 assemblies. Where IPR terms are in all assemblies they may have significantly different read counts. Most of these IPR terms and the related contigs exist in the other assemblies, but either due to incomplete assembly or simply fewer reads assembles they are not ranked with significance.

The remaining IPR terms in addition to not being including in all assemblies, are primarily protein functions that are not specific to tissue types. Unlike with the digestive gland, which requires the constant production of digestive enzymes to process food, the turnover of functional enzymes would be expected to be much lower for neural tissue. Comparison with other transcriptome analyses of Molluscan neural tissue would be the best scenario to deduce the expectations for neural tissue transcripts. However most analyses consider specific novel proteins and do not make any global assessment for the types of proteins found, making comparison difficult (Feng et al. 2009; Sadamoto et al. 2012).

4.6 Notable Transcripts in Neuronal Tissue

The presence of ribosomal proteins, actin, and cytochrome c oxidase appears more significant than in digestive gland; proteins such as these are also described in other mollusc neural EST libraries (Moroz et al. 2006). Whilst neural related proteins on the whole do not comprise a large section of the transcriptome, there are examples of proteins which can be related to neurological functions and do not exist in the digestive gland transcriptome such as G-Coupled Protein Receptors (IPR017452), FMRFamide peptides (IPR002544). In addition there are a number of proteins which contain notable properties that may make them of interest for further research, and are discussed further.

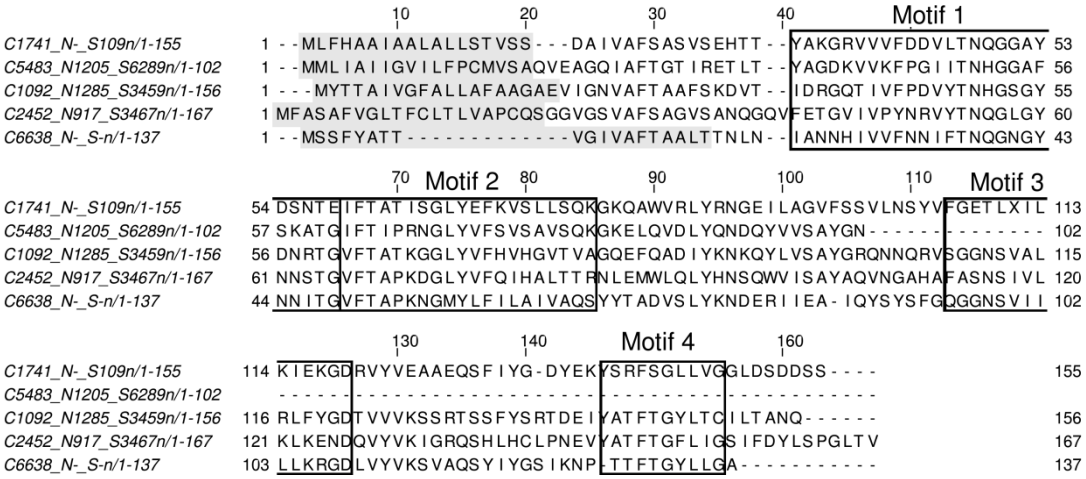


Figure 36 : Alignment of smaller TNF domain containing peptides

Alignment of 5 peptides from the D. reticulatum neural dataset, C5483;N1205;S6289n is only a partial peptide sequence with no stop codon prediction. Highlighted in light grey are the SignalP predictions for signal peptide regions. The boxed regions represent the 4 main motifs which characterise the Cq1 domain based on sprint accession PR00007 which was detected as part of the InterProScan homology analysis.

4.6.1 Tumour Necrosis Factor Like Proteins

8 predicted peptides contain a TNF-like domain (IPR006052). This domain is present in a range of proteins, which includes both Tumour necrosis factors and complement C1q. The TNF-like domain is present in both soluble proteins and proteins containing transmembrane domains. Polypeptides containing the TNF-like domain bind extracellularly to cysteine-rich receptors, and usually cause induction of intra-cellular apoptotic cascades. 5 of the *D. reticulatum* contigs, all of which are only available in the neural dataset, are incomplete sequences homologous to Complement C1q protein (IPR001073), see Figure 36. The IPR term describes C1q peptides as structurally similar to collagen VIII and X (Sellar, Blake, and Reid 1991; Muragaki et al. 1991) which assemble to form the C1q complex which activates the serum complement system. However C1q domain containing proteins (C1qDC) can be split into multiple subgroups not represented in the IPR term summary. C1q and

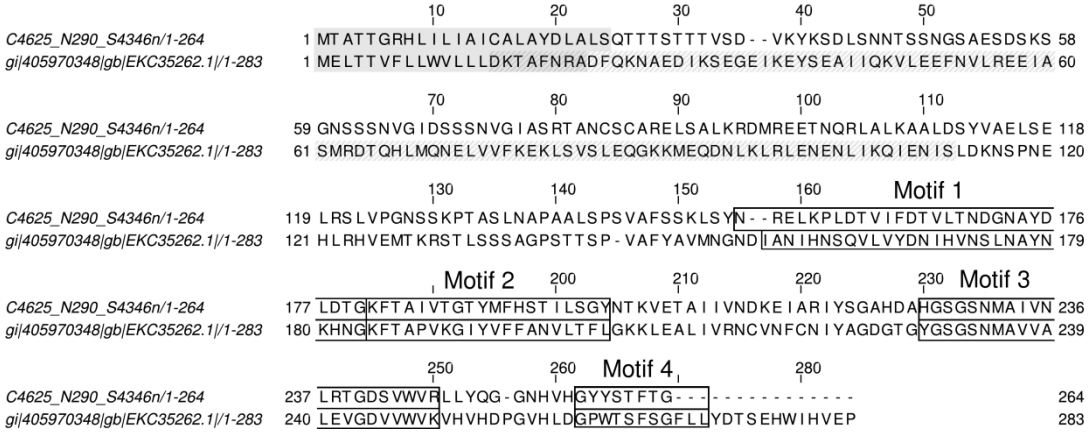


Figure 37 : Alignment of C1qDC peptide with homologue from C. gigas

Comparison of a top NCBI nr BLAST hit with D. reticulatum contig. Region highlighted in grey are predicted signal peptides, region in grey stripe for homologue is predicted TPR/MLP1/MLP2 conserved domain. The boxed regions represent the 4 main motifs which characterise the Cq1 domain based on sprint accession PR00007 which was detected as part of the InterProScan homology analysis.

C1q-like proteins are usually large (>2kb in length) and contain a partial or full collagen domain and a globular C1q domain. In contrast, globular head C1q proteins (ghC1q) contain a C1q domain with a short N-terminus region containing no motifs; they can either be secreted (sghC1q) or cellular (cghC1q) (Carland and Gerwick 2010). The predicted *D. reticulatum* peptides are short, with signal peptides, and as such all appear to be sghC1q proteins. In addition to the 5 smaller peptides a 6th one, shown in Figure 37, is present in the data, but has a significantly elongated N-terminal region. The top homologue from the NCBI dataset is of similar length and has a large number of conserved residues at the C-terminal C1q domain. However the presence of another TPR/MLP1/MLP2 domain detected through NCBI conserved domain (CDD) (Marchler-Bauer et al. 2011) in the homologue is not present in the *D. reticulatum* peptide.

Recently C1qDC proteins have begun to be characterised in invertebrate species, including c1q domain containing proteins in Mollusca. In Mediterranean

mussel *Mytilus galloprovincialis* 168 C1qDC transcripts were found, the vast majority of which contained signal peptide regions. A study of 8 transcripts indicated highly tissue specific expression levels, with 5 of 8 having low level constitutive expression in all but haemocytes, where expression was at a higher level (Gerdol et al. 2011). C1qDC peptides in *Chlamys farreri* were also assessed for immunological function, and were shown to interact with immunological related molecules as well as causing an increase in phagocytosis (Wang et al. 2012). The conclusion that they fulfill a pattern recognition role agrees with other studies of sghC1q proteins which can function as lectin-like molecules. With the C1q system being well characterised in mammals, most studies appear to be focused on C1qDCs correlating to immunological changes. However cerebellin proteins share very similar homology to other sghC1q proteins as well as with C1q complement peptides (Urade et al. 1991; Carland and Gerwick 2010). These proteins are around 190 residues long with a globular C1q domain at the C-terminus and a signal peptide, with significant expression levels in the Purkinje cells. These peptides form homohexamers and are secreted to the synaptic cleft were they function as adaptors to mediate interaction of variable pre- and post-synaptic cell-contacts (Eiberger and Schilling 2012). These proteins represent another potential function some of these *D. reticulatum* predicted peptides may have, particularly with their presence in the neural dataset. With increasing amounts of data about C1qDC peptides in multiple functions similar problems in predicting probable function occurs as with the C-type lectins discussed previously. These domains function as generic binding facilitators which evolutionarily have been adapted for a wide range of binding partners.

1 protein contains a TNF-domain but is linked specifically to the TNF ligand (IPR006052) rather than C1q. C-;pN48;S-n(C2550;S1758;N554d) exists only as a partial contig in the neural dataset, but the complete contig is available in the digestive gland, suggesting it may be ubiquitously expressed. Whilst relatively well studied in mammals, invertebrate homologues are less well characterised. Homology is primarily through position specific residue properties rather than specific conservation of amino acids, making orthologues appear less conserved. The *Drosophila melanogaster* homologue named Eiger has been shown to activate

Drosophila JNK pathway eliciting cell death (Igaki et al. 2002). The TNF proteins usually form trimeric complexes which bind TNF receptors (TNFR) on the cell surface. The factor can either be freely soluble or embedded in the cell membrane, The *D. reticulatum* contig appears to be the later, having a predicted trans-membrane region at the C-terminus, see Figure 38. In most examples of TNF ligands, the binding of the ligand to a cell surface TNFR elicits an apoptotic effect.

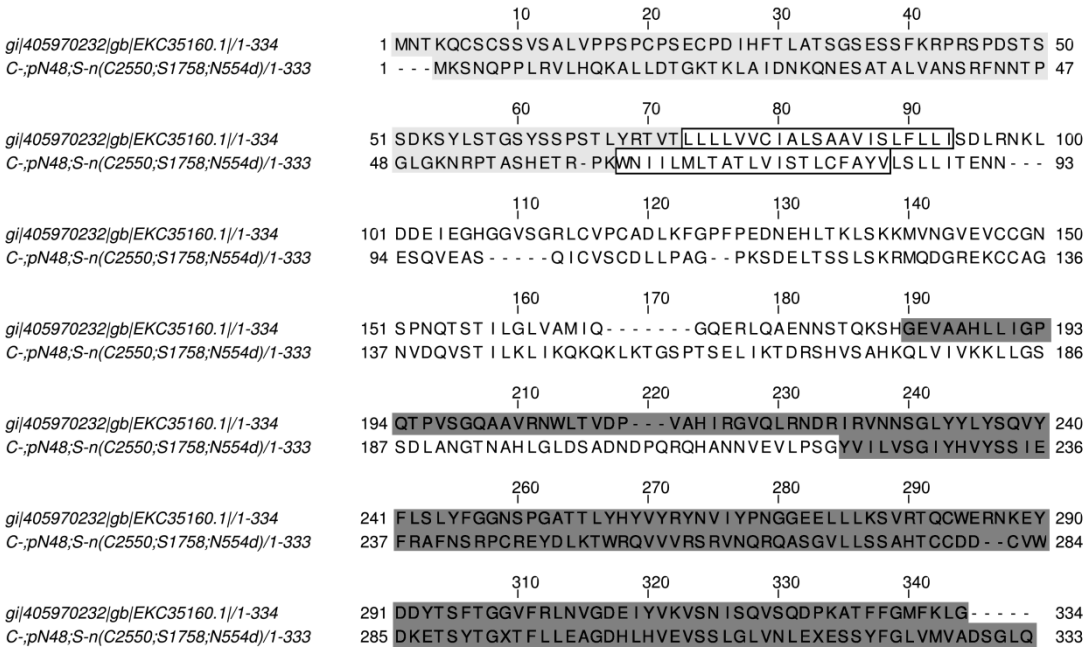


Figure 38 : Alignment of D. reticulatum TNF ligand like peptide with homologue

Highlighted regions: light grey, signal peptide, boxed, trans-membrane region, dark grey TNF ligand domain (IPR008983)

4.6.2 Vasopressin Like Peptide

IPR terms IPR000981 and IPR022423 represent the neurohypophysial hormone superfamily which contains vasopressin and oxytocin related hormone precursors. A predicted peptide of 164 residues shows strong homology to this group

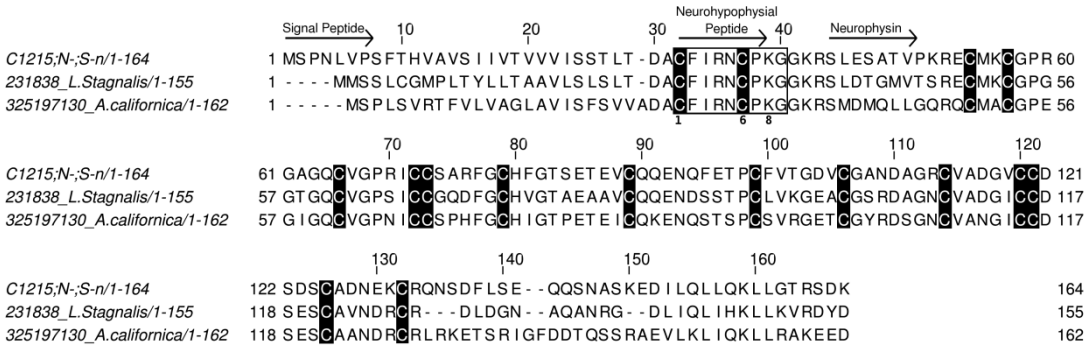


Figure 39 : Comparison of D. reticulatum conopressin like contig with homologues

The neurohypophysial peptide region is the best conserved between species and between the mollusc species shown we see they have identical sequence. All contigs also have a signal peptide and maintain cysteine positions, with the only variation in sequence length 5' or 3' of the first and last cysteine residues.

of proteins, in particular previously described Molluscan conopressin related precursor proteins. Unlike with previous protein groups described conservation amongst the vasopressin/oxytocin precursor superfamily is high and all peptides have well defined functionality. The hormone is translated as a precursor protein containing a signal region followed by the neurohypophysial peptide and then neurophysin. The peptide hormone itself contains 9 amino acid residues, with a disulphide bond between Cys1 and Cys6, and an amidated C-terminus. In vertebrates vasopressin hormones have both an antidiuretic and vasoconstrictive function to regulate water retention and blood pressure. Oxytocin mediates smooth muscle contraction, but due to similarity cross talk between functions of the two main types of neuropophysial hormone exist (Li et al. 2008).

Conservation between the *D. reticulatum* contig and the top 2 NCBI nr homologues is high with identical predicted neurohypophysial peptide hormone, and all 14 cysteine residues in the precursor, see Figure 39. The neurohypophysial site is the region which interacts with the neurophysin carrier region, allowing transport and aiding cleavage of the mature hormone (Kazmierkiewicz, Czaplewski, and

Ciarkowski 1997). Whilst the functional interaction of the neurohypophysial site is less well described in molluscs, conopressin G peptides contain the aromatic side chain amino acid residue at position 2 which, in vertebrates, is needed for interaction with neurophysin (Breslow 1979). The residue at position 8 divides the superfamily with conservation of basic lysine or arginine residues indicating a vasopressin related function rather than oxytocin (R. E. van Kesteren et al. 1992). However the unique feature commonly seen in the majority of mollusc vasopressin related sequences is the presence of a basic residue at position 4 giving a net charge of +3 compared to vasopressins (+2) and oxcytocins (+1) (Cruz et al. 1987).

The vasopressin like peptide present in *L. stagnalis*, which is the closest available homologue in the NCBI nr has an identical conserved neurohypophysial region and cysteine residues as the predicted *D. reticulatum* sequence. This peptide was shown to have functionality more similar to oxytocin despite having greater primary structure similarity to vasopressin related genes, with a synthetic version of Ile at residue 8 having functional equivalence to the native protein. This protein was localised to the CNS, the penis nerve, and the vas deferens and elicited muscle contraction when applied to vas deferens, indicating its role in reproductive behaviour (R. Van Kesteren et al. 1995). Whilst *L. stagnalis* is one of the few molluscs to have had the oxytocin/vasopressin superfamily characterised, in Cephalopoda the family has a functional role in long-term memory (Bardou et al. 2010). However the similarity between the Cephalopoda genes and *D. reticulatum* is not as great as *L. stagnalis*.

4.6.3 Spider Toxin-Like Protein

C1081;N102;S3902n and C-;N270;S3377n are contigs present only in the neuronal tissue and based on ESTScan predicted peptides match the term IPR004169, Spider toxin. BLAST results for these contigs give the highest similarity scores with insect homologues of around 86-125 amino acids in length. However none of these proteins have been characterised, but are described as 'Predicted spider toxins'. Both the unknown insect, *D. reticulatum* and spider toxins share a knottin-like structural motif, primarily conserved through the positions of disulphide-bond

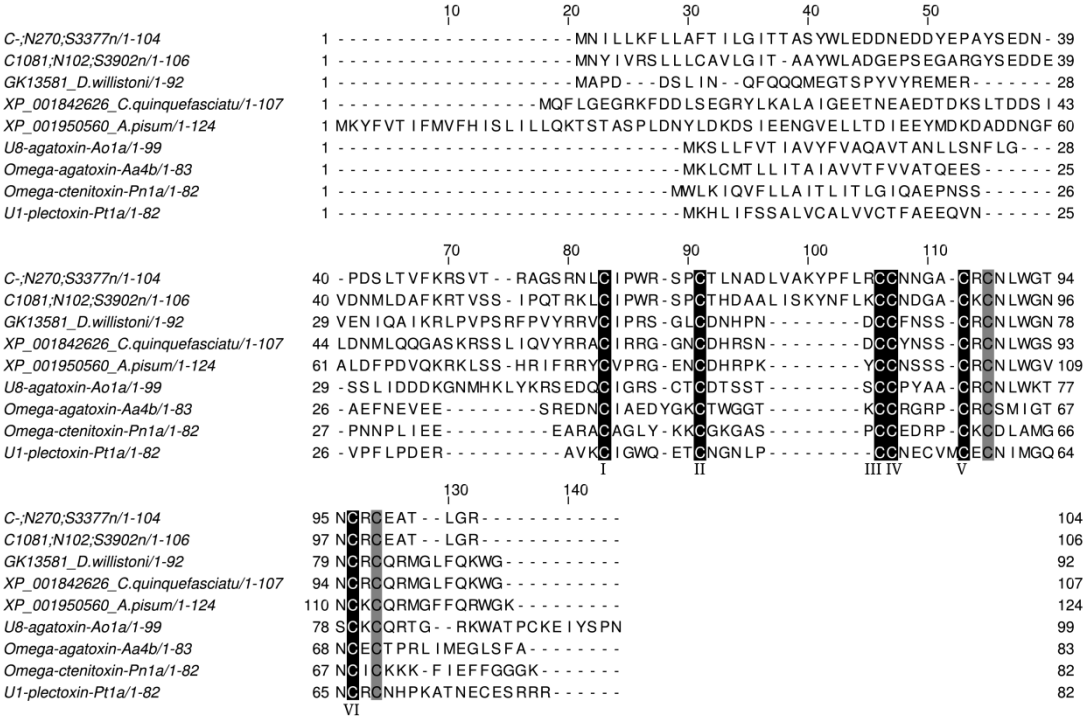


Figure 40 : Alignment of two D. reticulatum peptides with spider toxins and knottin containing insect peptides

Peptides included were those identified by BLAST searches against the NCBI nr, Arachnoserver or knottin databases. Cysteines in white on black labelled I-VI and predicted to form disulphide bridges between I–IV, II–V and III–VI to create the knottin structure. Cysteines highlighted in grey, may actually represent positions V and VI in some peptides, but without 3d protein structural analysis this is unclear. In some knottin domains V, VI and the greyed cysteine residues produce 2 disulphide bridges, totalling 4 in the knottin domain, giving extra stabilisation to the structure. C. quinquefasciatu and D. willistoni do not contain a 5' signal region; all other peptides have a 5' signal region, based on SignalP analysis.

forming cysteine residues. In many knottin domains the cysteine residues, labeled I to VI, and form 3 disulphide bridges between I–IV, II–V, III–VI, see Figure 40. In invertebrate knottin domains such as spider venoms and the chordate agouti-related domain, a fourth cysteine bridge is formed between cysteines near the V and VI

conserved cysteine residues (Gracy et al. 2008); these are present in the predicted *D. reticulatum* sequences.

Despite the homology to several classes of spider venom toxins the conserved cysteine residues have been identified in a number of other roles, with the proteins referred to as knottin mini-proteins. Knottins can describe any protein which contains the knotted cysteine-rich topology sharing the conserved disulphide bridge formations, most commonly playing inhibitory roles which in the case of spider and cone shell toxins block ion channel proteins (Gracy et al. 2008). The most similar homologous toxins to the *D. reticulatum* sequences can be found in the Arachnoserver database of spider venom toxins (Wood et al. 2009). Although many have had no biological activity characterised, those most similar to either *D. reticulatum* peptide are Ca_v channel blocking peptides. Excluding venom proteins, the online knottin database contains 2 groups of insect knottin protein families which are either mostly uncharacterised or are inhibitory proteins which have anti-microbial (Barbault et al. 2003) and anti-fungal functions. However using the built in BLAST search of the knottin database only presents either unknown insect peptides as with the nr database or spider-venom toxins as high quality homologues.

Another insect knottin-like domain has been shown to have similar homology to spider toxins, in *Musca domestica.* A novel phenol oxidase inhibitor (POI) includes a conserved cysteine-knot structure which homologous to a knottin domain (Daquinag et al. 1999). A phylogenetic tree of a variety of knottin-like domains is shown, Figure 41. The tree is limited to the cysteine knot region, from CysI-CysVI and suggests phenol oxidase is most similar to assassin bug venoms. With very different functions but similar homology this indicates the flexibility of the structural domain in a variety of functional roles. It is unlikely that the predicted *D. reticulatum* peptides have a similar function to the POI, which is only a 38 residue peptide with no additional cysteine residues to indicate a 4th disulphide bridge, present in the invertebrate ion-channel toxins and the *D. reticulatum* peptides.

Most spider venom toxins cause paralysis in a target prey organism with the venom peptides binding to ion channels. This binding is usually specific for a particular class of ion channels, with additional specificity discriminating between

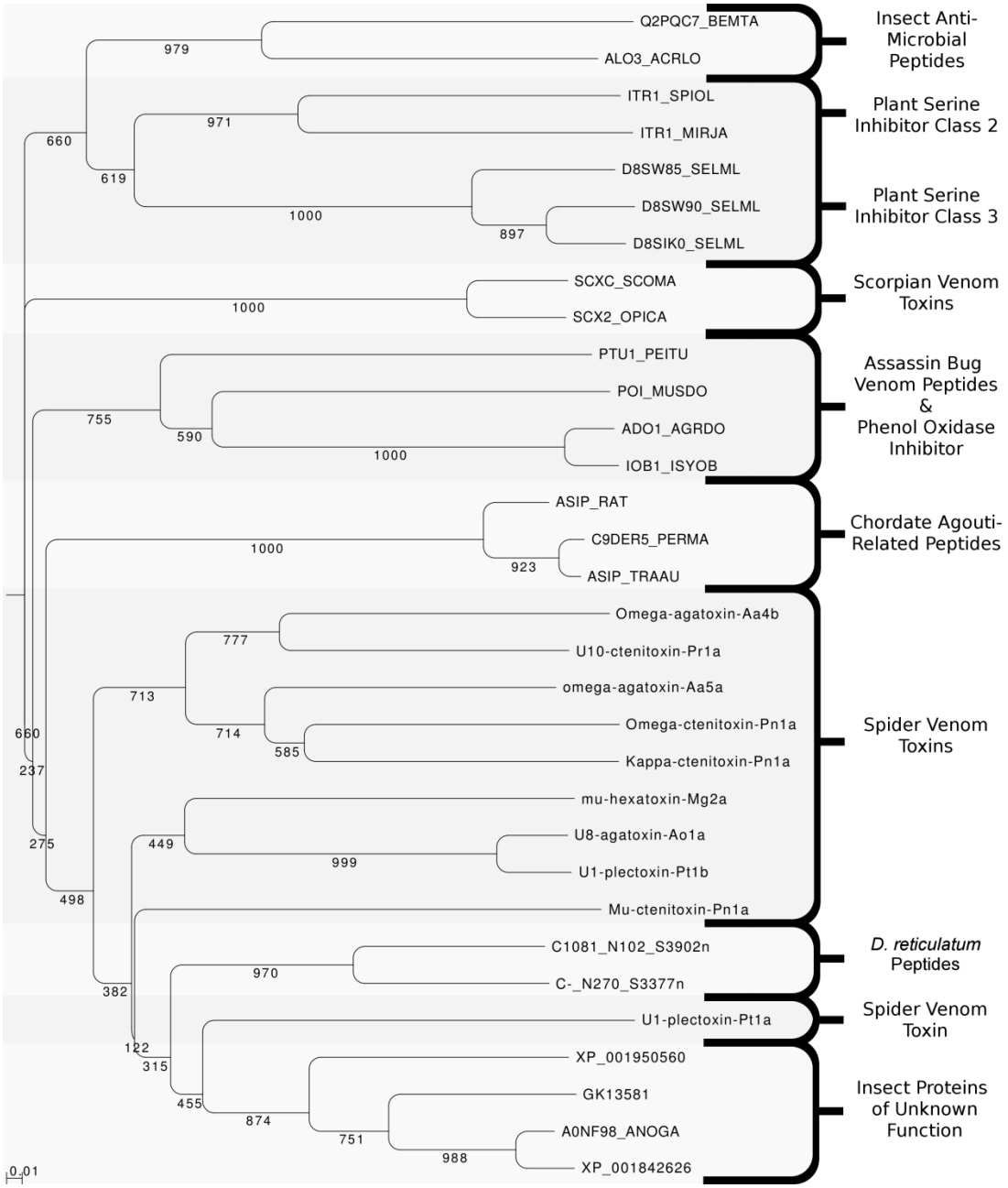


Figure 41 : Phylogenetic Tree based on the alignment of cysteine knot domains

Proteins of similar function appear to form clades together, with the primary exceptions being Phenol-Oxidase Inhibitor which groups with assassin bug proteins and U1-plectoxin-pt1a which sits within the D. reticulatum/insect clade. Agouti-related peptides, spiders toxins and unknown proteins group together and all share 2 cysteine residues at locations which in spider toxins have shown to form a 4th cysteine peptide bridge. Agouti-related peptides are hormones which bind to melanocortin receptors, which are a subgroup of G-coupled protein receptors.

Alignment was based on cysteine knot region from residues considered to be CysI to CysVI, bootstrap values are out of 1000. Sequences were taken either from BLAST hits results against NCBI nr database, Arachnoserver or the knottin database. Accessions are Uniprot, Arachnoserver IDs or NCBI accessions.

sub-classes of channels, and channels from different organisms. The strong homology of both *D. reticulatum* and insect peptides to spider-venom toxins may indicate a shared function in binding to ion channels, with the venom a structural mimic of native ion channel binding proteins. Alternatively the knottin domain may simply be an evolutionary building block, with only slight changes needed to allow for alternative functions. With the increasing number of examples of knottins in non-toxic roles, this peptide may represent an unknown family of knottins, which have utilised the knottin domain for a function which has no relation to ion channel binding. However the *D. reticulatum* peptides are present in neural tissue, with none found in digestive tissue, and interaction with endogenous ion channels must be considered as a possible function. Without experimental evidence no conclusion can be drawn, but the link between spider toxins and these peptides of unknown function found in both insects and *D. reticulatum* is intriguing. As such it may be of interest in further research as spider toxins are already being investigated as potential biopesticides (Windley et al. 2012). (Zhang et al. 2013). These *D. reticulatum* enzymes may be of interest for future research.

Chapter 5 | Application: Targeting *D. reticulatum*

With a wealth of data available from high throughput sequencing of *D. reticulatum* tissues, the next goal was to identify potential targets which could be used to develop molluscicidal agents. There are currently several main methods of transferring the molluscicidal agent to crop pest in common use. The first method of uptake is by combining the agent with a food bait such as the wheat pellets baits used for metaldehyde molluscicides. A second method is to apply the agent onto the target plant or plant seeds in a way which leads to consumption of both the plant and agent by the pest, as is done with spray application of most commercial insecticides. A third alternative is to incorporate the agent into the plant, primarily done with genetic modification and has shown to be very effective such as in *Bacillus thuringiensis* insecticidal toxin in cotton (Shelton et al. 2002).

The end goal for this study would be identifying targets which would cause mortality if their function were disrupted, either by down-regulation of the encoding gene with RNAi, or by direct effects on the protein product by a suitable protein or chemical inhibitor. The targets could be verified by injection or feeding assays with nucleic acids, proteins or small molecules produced *in vitro*. These assays would lead to new molluscicidal agents, which could be then be applied in crop defense in one of the previously described manners. Whilst oral toxicity would be the "end point" for potential molluscicides, the first step to identify any activity is best done with injection bioassay.

5.1 RNAi against Cathepsin L

Proteases in crop pests are a frequent target for endogenous defence mechanisms in plants, through the production of protein protease inhibitors and secondary metabolites such as tannins to inhibit insect proteases, thereby blocking protein digestion and the supply of nitrogen for nutrition. Production of protease inhibitor molecules is found as part of plant active defence (Ryan 1978; Van Dam et al. 2001) as well as part of constitutive, accumulated defence. The protein protease inhibitors produced by plants show some specificity towards different herbivorous

pests, showing activity towards specific protease mechanistic classes; different orders of insect pests show differing digestive biochemistry as an adaptation to this defence. Down regulation of proteases in insects should enable endogenous plant defence based on protease inhibitors to work more effectively. RNA interference studies which utilised dsRNA with sequence specific to Cathepsin L in the pea aphid, *Acyrthosiphon pisum*, showed promising results, with gene expression knock down within 1 day post-injection (Jaubert-Possamai et al. 2007) after injection into the body cavity, and a maximal decrease in gene expression of 40%. This RNAi study had multiple benefits, both identifying if an RNAi effect could be elicited, as well as indicating whether this Cathepsin L was necessary for the organism's survival. The transcriptomic analysis showed that a cDNA encoding Cathepsin L corresponded to a major transcript in digestive gland, indicating its importance to *D. reticulatum* as a major digestive protease, and therefore it presented a promising target for a gene knock-down study. A phenotype in the organism resulting from successful down-regulation would also suggest the likelihood that cysteine protease inhibitors could be used as a pesticide against the organism.

5.1.1 dsRNA Production and Injection Assay for RNAi Effects

The transcriptomic data available for *D. reticulatum* avoids the need to isolate cysteine protease genes via production of degenerate primers, followed by cloning, sequencing and potential 5' and 3' RACE, as was initially done with the ferritin gene in this study. Instead, a transcript encoding the major cysteine protease in digestive gland was identified by database search. Constructs were designed for production of two dsRNA fragments, corresponding to 5' and 3' regions of the Cathepsin L transcript of *D. reticulatum* (5'Cath & 3'Cath), of size 415bp & 419bp respectively. The fragments were amplified with PCR from digestive gland tissue first strand cDNA, and their size was confirmed by visualisation with 0.8% Agarose gel, see Figure 42. PCR fragments were isolated and purified via QIAquick Gel Extraction Kit [Qiagen] then cloned into an equivalently digested pLitmus28i vector. RNAi Cathepsin L constructs were transformed into *E. coli* TOP10 strain and sequence checked through DNA sequencing.

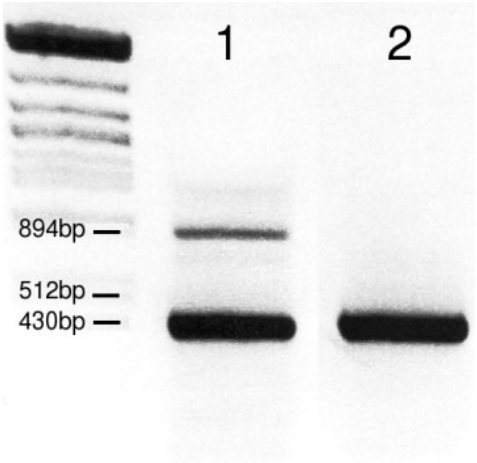


Figure 42 : PCR products Cathepsin 5' and 3' Fragments for RNAi Constructs

Lane 1 includes PCR product using the 5' Cath L RNAi construct and lane 2 includes the 3' Cath L RNAi primers. Fragments of around 410-420bp seen on this gel image were isolated by gel extraction and utilised for molecular cloning of the full RNAi construct. Lane 1 appears to have a small amount of secondary product, but this was excluded via gel extraction and the sequence of both fragments confirmed with sequencing of final clones.

A control dsRNA was used for the RNAi experiment; in this case a sequence encoding bacterial neomycin phosphotransferase II which gives resistance to the antibiotic kanamycin was used. Kanamycin is a commonly used antibiotic; the 795bp resistance gene is used as a control due the very low probability of it being present in the *D. reticulatum* genome. As such it can provide a positive control to demonstrate that it is the specificity of the dsRNA sequence that causes any gene regulatory effect observed rather than the presence of any dsRNA. The gene sequence is inserted into pLitmus28i vector which can be transcribed in parallel with the target gene constructs. To demonstrate that injected media effectively infiltrates the Molluscan haemolymph, test injections were conducted with PBS buffer dyed with food colouring. The dye progresses throughout the organism within a minute of injection

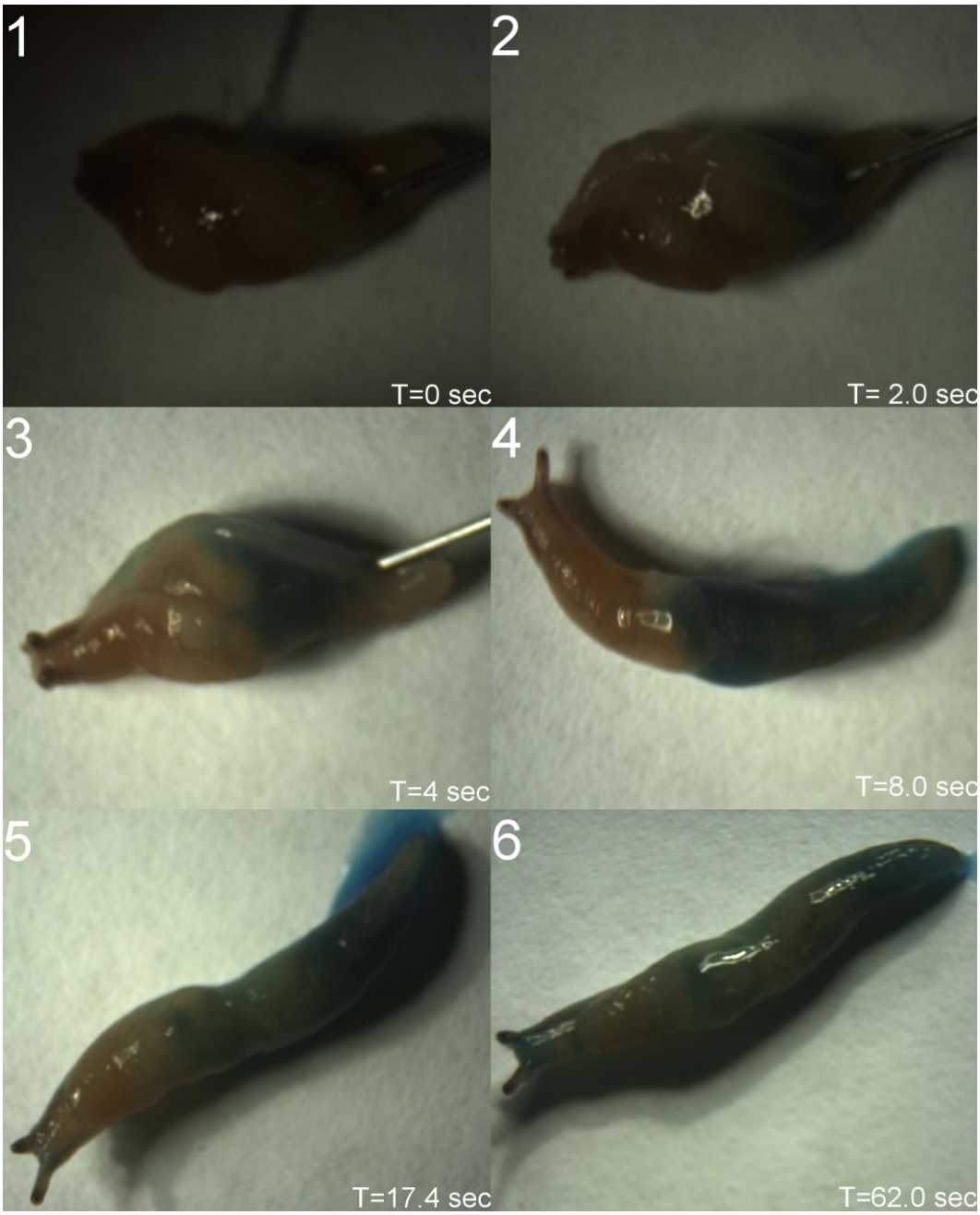


Figure 43 : Dye injection assay

Frames 1-6 represent stills from a video recording of a dye injection assay. T= represents time and begins at 0 seconds in frame 1 at the point of injection. Frame 2 and 3 shows the infiltration of 10µl dyed PBS as it is injected into the organism. After 4 seconds the injection needle is removed and the resulting spread of dye monitored. Frames 4 & 5 show the progression and indicate dye in the trail of the slug. At around Frame 6 1 minute after injection, the maximum amount of infiltration is seen with dye observable at the head.

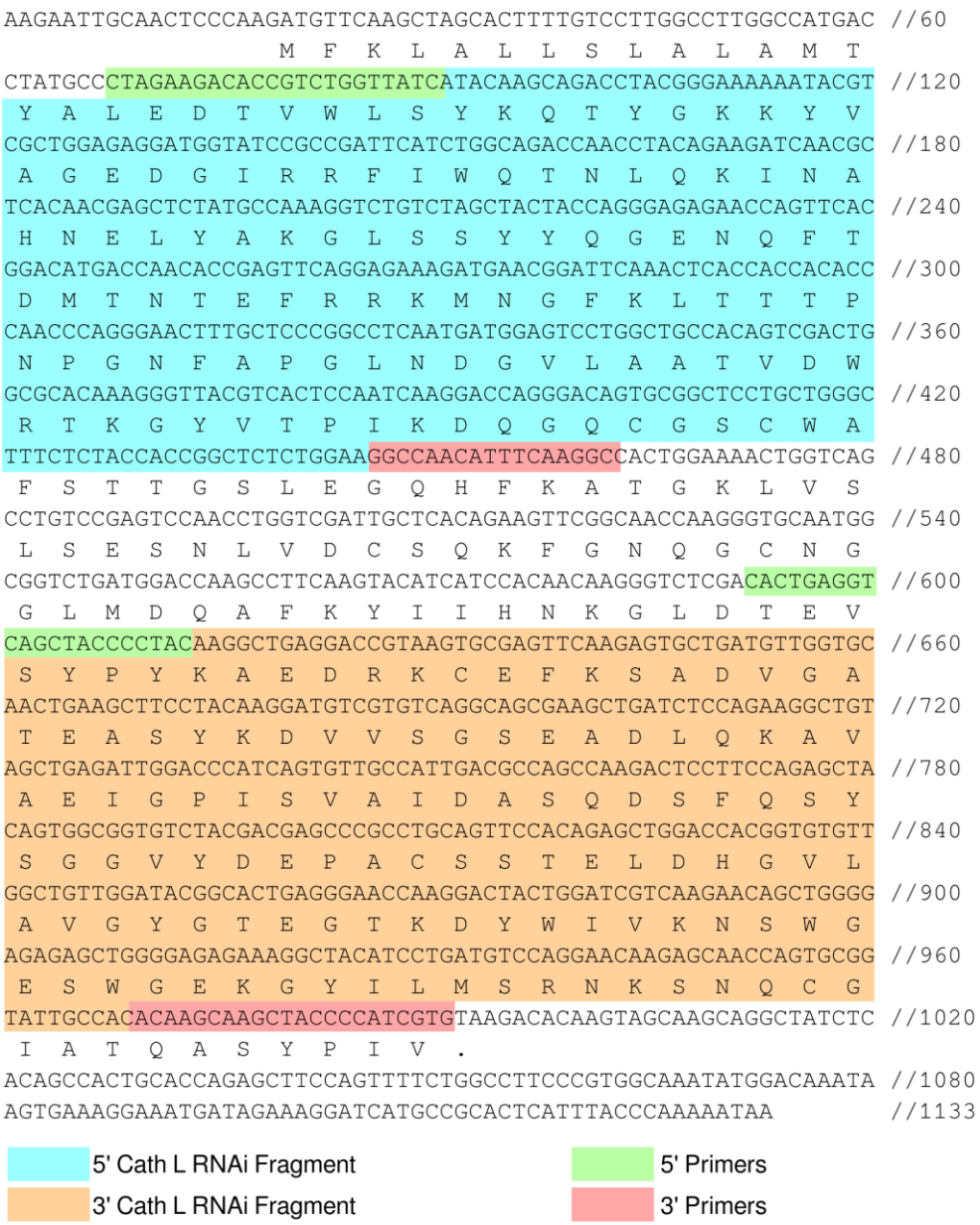


Figure 44 : Cathepsin L RNAi fragments

The figure shows the cDNA sequence and predicted translated region for a cathepsin L homologue previously identified. 5' and 3' Cathepsin fragments for insertion into RNAi vector are highlighted in blue and orange respectively. Regions primed by 5' and 3' primers for each fragment's initial amplification from cDNA are highlighted in green and red respectively.

and appeared to cause no harmful effects. Notably dyed mucus was exuded by the organism suggesting an increasing amount of injection material may be lost over time, see Figure 43.

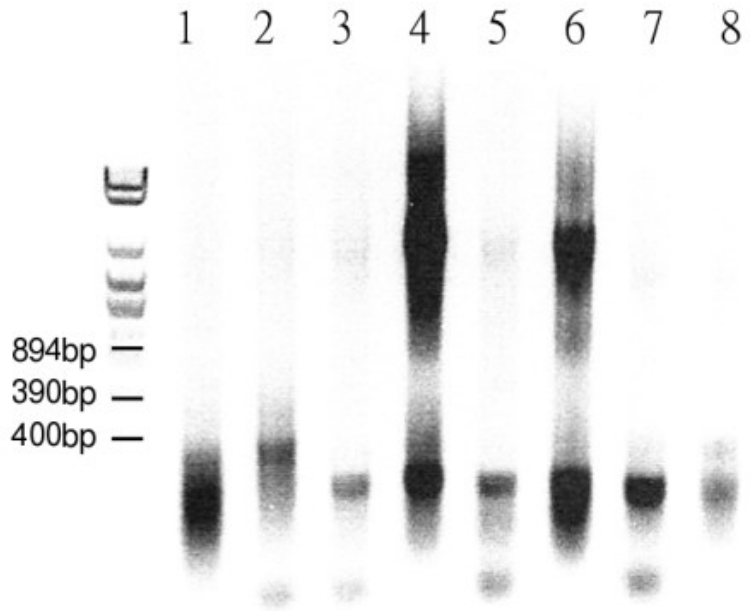


Figure 45 : ssRNA for Kanamycin, 5' & 3' Cath and Apin RNAi experiments

Lanes 1 & 2 represent Kanamycin 5' and 3' ssRNA. Lanes 3, 4, 5, 6 represent 5' and 3' ssRNA for 5'Cath and 3' Cath RNAi construct. Lanes 4 & 5 appear to contain higher molecular weight ssRNA, possibly due to incomplete linearization with 3' restriction enzyme. Lanes 7 & 8 contain ssRNA produced from APIN RNAi construct. ssRNA often appears smeared on non-denaturing Agarose gel electrophoresis. Sizes appear correct, but were confirmed by annealing and assessing dsRNA sizes. High molecular weight ssRNA was no longer observable in dsRNA gels after annealing.

Primers were designed to PCR fragments from 5' and 3' regions of the predicted Cathepsin L transcript for insertion into RNAi transcription vectors. The regions used for 5' and 3' Cath RNAi constructs are shown in Figure 42. 5' Cath, 3'Cath and Kanamycin constructs were used to produce ssRNA though *in vitro* transcription, see Figure 43 which was then quantified using NanoDrop^{TM} 1000

spectrophotometer (Thermo-Fisher) and image estimates from 0.8% Agarose gel electrophoresis and annealed together to generate dsRNA. dsRNA products were visualised on 0.8% Agarose gel to confirm size, see Figure 46, then quantified using a NanoDrop™ 1000 spectrophotometer (Thermo-Fisher). 5' Cath, 3'Cath, Kanamycin dsRNA were diluted in PBS to a concentration 2µg/µl and 5µl was injected into 5 organisms for each respective dsRNA (15 organisms total). Alongside this, 5 organisms were 'sham' injected with PBS only. Each group of 5 were kept separate on diet and monitored periodically. After 48 hours no phenotype was observed, digestive glands were dissected from individuals and flash frozen and prepared for qPCR analysis.

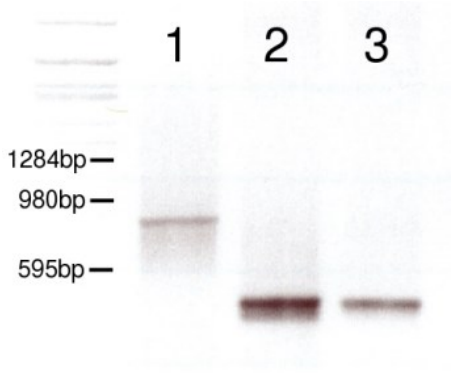


Figure 46 : in-vitro transcribed dsRNA of Kanamycin and Cathepsin L fragments

Lane 1 contains Kanamycin dsRNA, Lane 2 and 3 contain the 5' and 3' Cathepsin gene fragment dsRNA. There is a small amount of smearing, but the clear bands at expected positions indicate dsRNA is not degraded.

5.1.2 Quantitative qPCR on Cathepsin L RNAi injections

Dissected frozen digestive glands were ground in liquid nitrogen and total RNA was extracted. Purified total RNA was used as a template to synthesise first strand cDNA. The resulting cDNA was used in a comparative Ct qPCR study of the cathepsin L gene expression under different treatments; the results were normalised to GAPDH levels as shown in Figure 47. The data did not indicate a consistent RNAi

Compare Sample Groups	p-Value (Independent T-Test)
Control vs. Kanamycin	0.74
Control vs. 5’ Cath L	0.25
Control vs. 3’ Cath L	0.16
Kanamycin vs. 5’ Cath L	0.45
Kanamycin vs. 3' Cath L	0.12

Table 21 : T-test values of Cathepsin L RNAi experiment

Table includes p-value results calculated by comparing 5 RQ values of each group against each other using an independent T-Test function. In all cases no group can be considered to have a statistical significant difference with no p values below 0.05.

effect produced by injecting cathepsin dsRNA; whilst the overall average for the 3' RNAi fragment of cathepsin L represents a 38% drop in expression, similar to described in (Jaubert-Possamai et al. 2007), no drop in expression was produced by the 5' RNAi fragment (see Figure 48). The internal deviation of the samples such as the 40% increase seen in the 5' Fragment results indicate the levels could equally be explained by normal variation within the sample groups.

The observed variance could likely be due to the nature of the organism, having more atypical feeding habits, as compared to insects. Whilst a strong positive result would have been effective proof of a gene knock down effect, a negative result and such large variance, calls into question a number of things. The choice of a gene which has a high variance normally may have been a mistake in an organism where no previous RNA interference had been proven. The use of GAPDH as a standard for normalising gene expression, whilst commonplace in insect species, may not be as valid in molluscs. The choice of GAPDH is based on the role of as an enzyme in the central metabolism, where it is designated a "housekeeping gene" (HKG) and is assumed not to vary in expression between different tissue types and external conditions. Other genes have been suggested as HKG for normalisation in qPCR

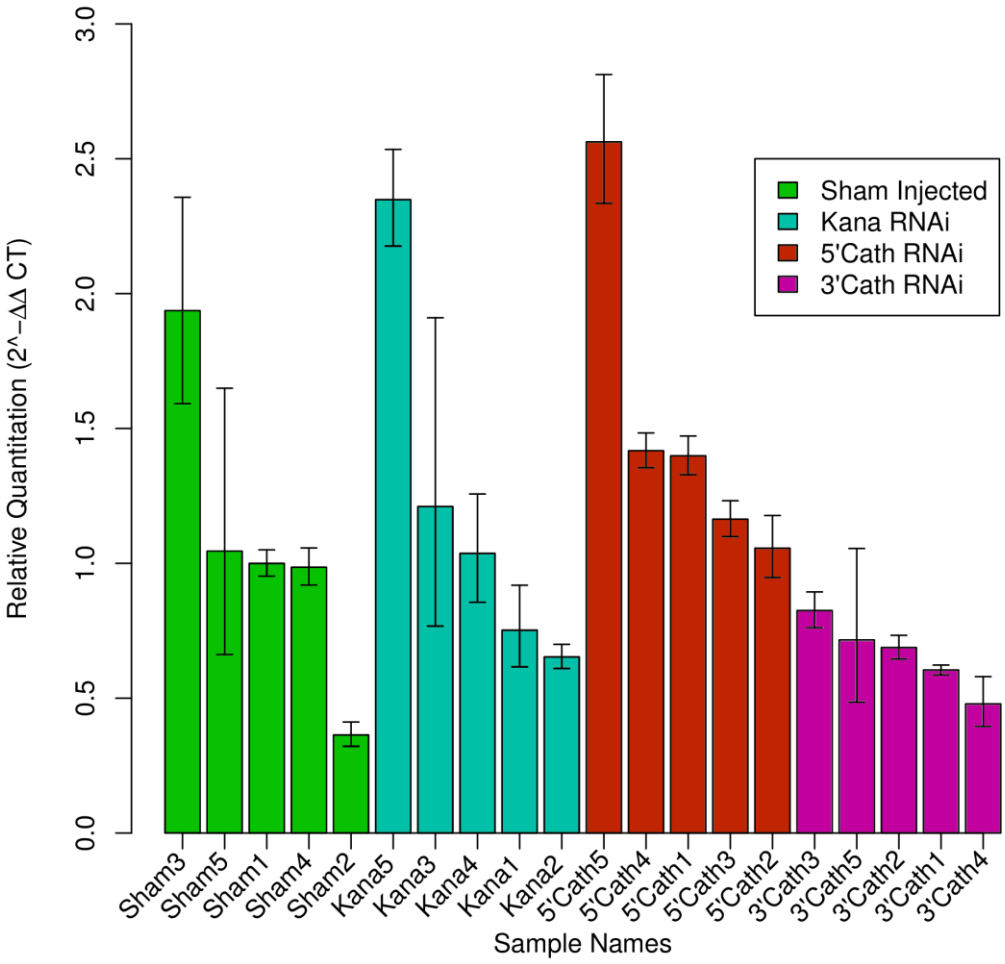


Figure 47 : Cathepsin Levels for individual organisms in RNAi Study

Figure shows the relative quantitation of Cathepsin L gene, against GAPDH. Sham Injected and Kanamycin represent controls, each include variance of 0.5-2.5. 5'Cath and 3'Cath samples names represent individuals injected with either of the 5' or 3' dsRNA fragments. Overall no significant gene knock down effect can be seen with groups not being significantly different, see Table 21. Error bars shown represent 1 standard deviation based on 3 technical replicates for each sample.

experiments. Previous Molluscan RNAi studies seem erratic in HKG choice, but they often backed data up with phenotypes or protein assays.

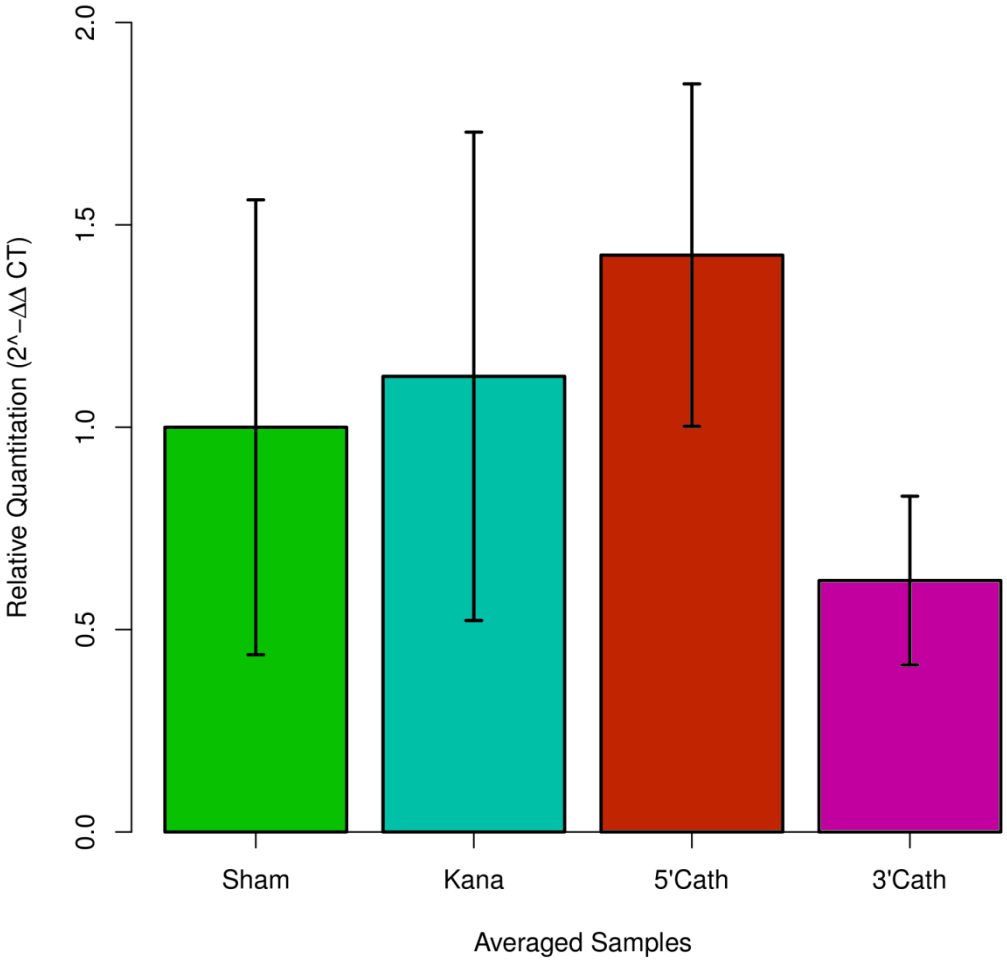


Figure 48 : Average RQ values for RNAi against Cathepsin L

Figure shows the averaged values of individuals injected with dsRNA for Kanamycin, 5' Cathepsin L Fragment, 3' Cathepsin L fragment or sham injections. The resulting data indicated relatively poor consistency of expression in individuals. Whilst 3'Cathepsin L fragment shows a decrease it is equal to the increases seen in other data and no group is statistically significant, see Table 21. Error bars shown represent 1 standard deviation based on the 5 biological replicates used to generate each average value.

5.2 Assay of Normalisation Standards for qPCR (HKGs) for *D. reticulatum*

The failure to show a clear RNAi effect in the cathepsin L down-regulation experiment called into question the validity of GAPDH as a house keeping gene; variability in expression of this gene was one potential explanation for the high level

of variability in gene expression observed in the qPCR assays. Research on another mollusc, *Octopus vulgaris,* demonstrated the use of comparing multiple HKGs to

HKG	Delta CT	Best-keeper	Norm-finder	GeNorm	Overall Rank
Actin	0.74	0.35	0.3	0.38	1.19
EFI-A	0.76	0.3	0.33	0.38	1.41
GAPDH	0.88	0.62	0.55	0.55	3.46
Tubulin	0.99	0.39	0.86	0.75	3.46
Ubiquitin	1.13	0.97	1.03	0.9	5

Table 22: Table of 5 HKG analysed for stability using qPCR Ct value data

The table shows the stability scores produced by the relevant HKG stability tools, lower values represent greater stability. The final overall value represents an aggregate rank; with 5 represent lowest stability score in all analyses and 1 highest. In this case we see Actin and EF1-A as the best pair produced by GeNorm which returns a top pair with a joint score. Overall the stability difference between EF1-A and Actin is very similar within 0.1 in all analyses. GAPDH has the next highest stability although Bestkeeper indicates poorer stability than tubulin. Overall both tubulin and ubiquitin are considered the poorest, and may relate to the high Ct values seen suggesting poorer qPCR efficiency.

select the one with lowest overall variance (Sirakov et al. 2009) via either pairwise comparison (Vandesompele et al. 2002; Andersen, Jensen, and Ørntoft 2004) or comparison to artificial values generated from the total data (Pfaffl et al. 2004). Using the transcriptomic data set, primers for quantitative PCR were designed based on sequences which were identified as EF1-A (C1490;N398;S553d), Actin

Digest & Whole	Delta CT	Best-keeper	Norm-finder	GeNorm	Overall Rank
Actin	1.39	1.27	0.65	1.3	1.19
EF1-A	1.67	1.15	1.36	1.3	1.41
GAPDH	1.76	2.03	1.52	1.61	3

Table 23: HKG analysis of 3 HKG with digestive gland and whole tissue qPCR data

Table shows 4 analyses of Ct values produced from 3 HKG using 11 digestive gland tissues and 4 whole tissue samples. Actin and EF1-A are considered the most stable pair by GeNorm and EF1-A with considered more stable by Bestkeeper but less stable than Actin by Normfinder and Delta CT. The combination of whole and digestive gland leads to lowest stability when using GAPDH.

(C2471;N689;S1995d), GAPDH (C2454;N37;S199d), Beta-Tubulin (C2968d), and Ubiquitin (C408d). RNA was extracted from digestive glands of 5 organisms and cDNA synthesised. qPCR Analysis was conducted as described (Sirakov et al. 2009) using the comparative ΔCt method for data analysis. The resulting analysis suggested Actin, followed closely by EF1-A were had the highest stability, see Table 22. However the Ct values for Tubulin and Ubiquitin were particularly high, and may represent poor qPCR efficiency rather than low stability.

The top 3 qPCR HKG were further investigated. Six additional digestive gland tissue and 4 whole tissue samples were used for qPCR analysis, and their Ct values were used for further comparison. Interestingly, comparison of the 3 best ranked genes remains the same even when the additional 6 digestive gland samples are included (see Table 23). However inclusion of 4 whole tissue samples, with digestive gland data, alters the ranking with Actin as the best HKG. When whole tissue sample Ct values are used on their own though GAPDH is the best HKG, see Table 24. When considering the raw Ct values we see that GAPDH is notably higher than the 2 other genes but very similar, despite the square increase that Ct represents, see Table 25. In addition all 4 analyses indicate Actin as the worst HKG in whole

Whole Tissue	Delta CT	Best-keeper	Norm-finder	GeNorm	Overall Rank
GAPDH	1.17	0.29	0.46	0.92	1
EF1-A	1.47	0.52	1.21	0.92	1.68
Actin	1.73	1	1.63	1.46	3

Table 24: HKG analysis of 3 HKG with whole tissue qPCR data

Table shows the same 4 whole tissue samples from whole the previous comparison but analysed separately. We see a shift between GAPDH and Actin switching places in rank. Overall EF1-A has similar stability values for each individual analysis as previous and does not appear to be notably different for whole tissue.

tissue. As such depending on the tissue source and data input anyone of these genes could be considered most 'stable'.

Overall, variability in gene expression between individuals in *D. reticulatum* populations appears to be greater than insect equivalents. Unlike insects, which go through very pronounced developmental stages, *D. reticulatum* live longer, with variable growth and less regular feeding and activity. Previous work shows slugs hatched together have asymmetrical growth leading to slugs of equivalent age varying up to 100-fold in mass, and through inference slugs of equivalent mass can be of dramatically different age. These factors combined with a much smaller body of works with previous Molluscan qPCR make designing qPCR experiments more difficult. This analysis identifies weaknesses with the use of qPCR with this organism. Future quantitation may need to pool multiple organisms, and use multiple reference genes in order to better validate results. Despite this both analysis using methods previously described by Sirakov et al and basic observation of Ct values suggests EF1-A to be the best all round HKG, consistently having high stability of gene expression regardless of tissue used. After this finding, further qPCR experiments used EF1-A as the HKG of choice.

Tissue	Actin	SD	EF1-A	SD	GAPDH	SD
Digestive Gland (1)	**18.72**	0.21	**19.02**	0.04	**21.21**	0.06
Digestive Gland (1)	**19.48**	0.21	**19.41**	0.15	**22.03**	0.10
Digestive Gland (1)	**19.12**	0.10	**19.94**	0.01	**22.77**	0.11
Digestive Gland (1)	**18.89**	0.05	**19.13**	0.18	**23.39**	0.05
Digestive Gland (1)	**19.67**	0.12	**19.67**	0.03	**22.57**	0.12
Digestive Gland (2)	**20.46**	0.14	**21.00**	0.03	**23.78**	0.08
Digestive Gland (2)	**22.39**	0.34	**23.03**	0.02	**25.55**	0.08
Digestive Gland (2)	**21.56**	0.06	**22.49**	0.13	**24.35**	0.19
Digestive Gland (2)	**21.09**	0.07	**22.54**	0.34	**25.92**	0.06
Digestive Gland (2)	**21.85**	0.07	**22.82**	0.06	**26.2**	0.07
Digestive Gland (2)	**20.63**	0.20	**21.52**	0.26	**24.78**	0.06
Whole	**20.24**	1.46	**22.11**	0.11	**28.2**	0.02
Whole	**22.54**	0.03	**20.39**	0.31	**28.1**	0.07
Whole	**23.13**	0.33	**20.72**	0.14	**28.73**	0.06
Whole	**23.03**	0.01	**21.07**	0.47	**27.65**	0.11

Table 25: Table showing Ct values of 3 HKG in digestive gland and whole tissue

Digestive gland (1) and (2) represent two 'batches' of samples prepared and quantitated with qPCR. Whilst whole tissue Ct values are on the whole higher, relative to one another GAPDH appears significantly higher. The same starting RNA and cDNA was used for all qPCR results. The SD value represents 1 standard deviation of the Ct of the 3 technical replicates used for each sample.

5.3 Down Regulation of Expression of Apoptosis Inhibitor by RNAi

Apoptosis Inhibitor (APIN) proteins are a group of proteins characterised by the presence of a baculovirus inhibitor of apoptosis protein repeat (BIR) domain. They suppress apoptosis via interacting with and inhibiting the caspases which cause the proteolytic cascade in the apoptotic process. Down-regulation of APINs in insects using RNA interference has been shown to induce apoptosis in cells (Liang et

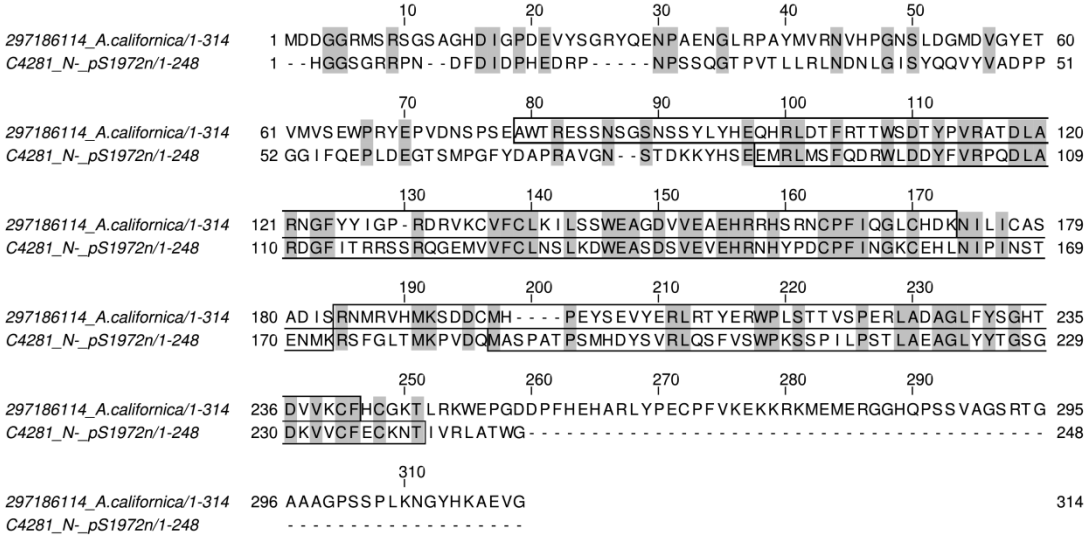


Figure 49 : Comparison of D. reticulatum contig with A. californica apoptosis inhibitor

A. californica apoptosis inhibitor was the top match for this contig from the NCBI nr database, with matching residues highlighted. Both sequences contain 2 BIR domains shown by boxed regions, giving strong evidence for the D. reticulatum contig having a similar function. The D. reticulatum contig appears to not be a complete CDS, but target RNAi insert size is 300-400bp so a complete sequence is not necessary for this experiment.

al. 2012; He et al. 2012). The potential to stimulate large scale apoptosis in an organism maybe one of the best ways of eliciting a phenotype. As such *D. reticulatum* contigs containing BIR domains were identified, C4281d;N-;pS1972n was selected as it had a homologous region large enough to design a RNA construct, see Figure 50. Although the main aim of the study was to cause a phenotype in the organism, using a dsRNA fragment derived from this contig, quantitative analysis of gene down-regulation was also required. Primers for qPCR which would amplify a region outside the dsRNA fragment were designed but produced poor results when used in trial experiments. This was either due to lower levels of transcript or the

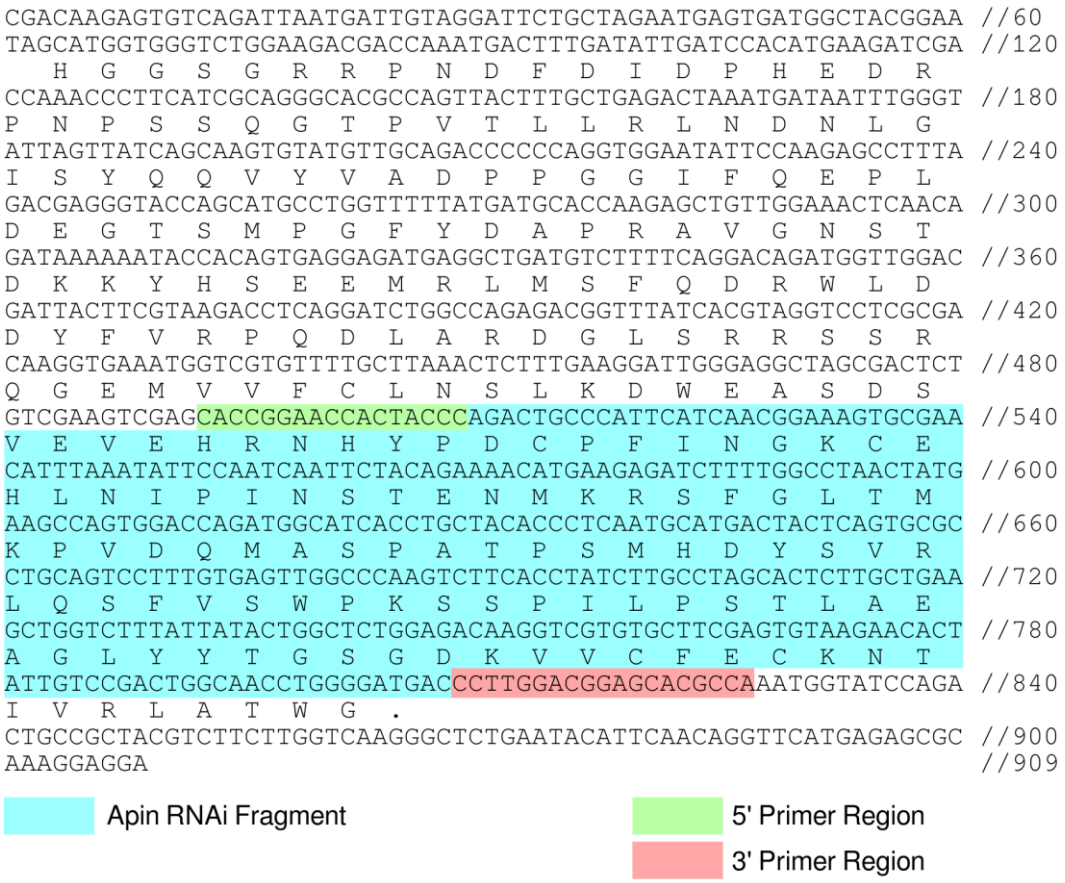


Figure 50 : Apoptosis Inhibitor RNAi fragment

The figure shows the cDNA sequence and predicted translated region for a apoptosis inhibitor homologue. The fragment for insertion into RNAi vector is highlighted in blue and regions primed by 5' and 3' primers for initial amplification from cDNA are highlighted in green and red respectively.

regions outside the RNAi fragment, from where primers were designed, being a poorer quality sequence with more errors. Despite this the dsRNA fragment was still made and injected to ascertain potential for phenotypic effect.

5.3.1 RNAi Construct & Assay

Primers were designed used to amplify a region of 342bp using whole tissue first strand cDNA as template, see Figure 51. The 342bp fragment was amplified via PCR, size assessed via Agarose gel electrophoresis, Figure 51 (A). The PCR fragment was then cloned into the pJet 1.2 which was used as a transcription vector

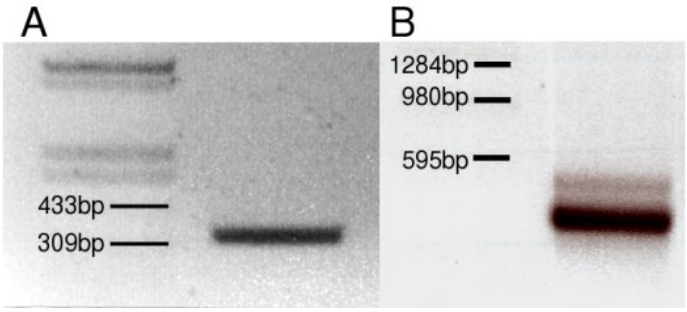


Figure 51 : D. reticulatum Apoptosis Inhibitor PCR and dsRNA electrophoresis gels

A) 0.8% Agarose gel electrophoresis of Apoptosis inhibitor PCR product which has clear band that matches the predicted size of 342bp. This was gel extracted and cloned into a transcription vector.

B) 0.8% Agarose gel electrophoresis checking the quality of the Transcribed apoptosis fragment dsRNA. There is limited low molecular weight smearing suggesting the dsRNA is of relatively good quality.

and checked by sequencing 5' and 3' of the inserted region. ssRNA was synthesised, see Figure 43, via *in vitro* transcription as described before, alongside kanamycin control and annealed to produced dsRNA. dsRNA was assessed by gel electrophoresis, see Figure 51 (B). 20µg APIN dsRNA was injected per organism, in a total of 10 organisms, which was mirrored with Kanamycin dsRNA and buffer only sham injections, so that 30 organisms in total were used in the experiment. Individuals were fed on lettuce and wheatgerm diet at 10°C and monitored for 2 weeks for phenotype changes. After 2 weeks no phenotype change was observed. 2 individuals in the group had died (1 sham, 1 APIN), but this not unusual for individuals kept in the Lab, and was not considered a phenotype change. Whilst more assays at differing concentrations could be done, due to the lack of results and availability of organisms alternative studies were considered. It should be noted that studies cited for insects were primarily in cells and embryos rather than whole organisms. Work currently being done indicates apoptosis can be induced in adult insects, but effects on phenotype can be most detrimental during developmental changes such as pupation (Pyati, 2012, Unpublished). Injecting adult molluscs may

```
GTTCAGATCGCCGGTAGAACTGATACATTAATAAAATGGCTGAACTGAAAGTTGGAATTA //60
                                    M  A  E  L  K  V  G  I  N
ACGGATTTGGGCGCATTGGTCGTCTGACCCTGCGTGCTGCGCTCCAGAAAAACGTTAATG //120
  G  F  G  R  I  G  R  L  T  L  R  A  A  L  Q  K  N  V  N  V
TTGTTGCAGTCAATGATCCTTTCATTAGTCTGGAGTACATGGTCTACATGTTTAAGTATG //180
  V  A  V  N  D  P  F  I  S  L  E  Y  M  V  Y  M  F  K  Y  D
ATTCTACACACGGCCGCTATAAGGGTAAAGTGGCAGAGAACAATGGCAAACTTGAGATCG //240
  S  T  H  G  R  Y  K  G  K  V  A  E  N  N  G  K  L  E  I  D
ATGGGCACCTCATCACAGTCTTTGCTGAGCGAGATCCAGCTGCCATTAACTGGAAGTCTG //300
  G  H  L  I  T  V  F  A  E  R  D  P  A  A  I  N  W  K  S  A
CTGGGGCCAACTATGTGGTAGAGTCAACTGGAGTGTTCACAACCGCTGACAAAGCAAATG //360
  G  A  N  Y  V  V  E  S  T  G  V  F  T  T  A  D  K  A  N  V
TGCACATAAAAAGTGGAGGTGCTAGCAAGGTTGTAATCTCTGCACCATCTGCTGATGCTC //420
  H  I  K  S  G  G  A  S  K  V  V  I  S  A  P  S  A  D  A  P
CAATGTTTGTGATTGGTGTAAACGACGACAAATACACGAAGGACATGACTGTGGTCAGTA //480
  M  F  V  I  G  V  N  D  D  K  Y  T  K  D  M  T  V  V  S  N
ATGCTTCCTGCACAACTAACTGTCTGGCCCCCCTGGTCAAGGTCATCAATGACAATTTTG //540
  A  S  C  T  T  N  C  L  A  P  L  V  K  V  I  N  D  N  F  G
GCATCGTTGAGGGTTTGATGACAACCGTACATGCTACCACCGCCACACAGAAGACTGTAG //600
  I  V  E  G  L  M  T  T  V  H  A  T  T  A  T  Q  K  T  V  D
ATGGACCCAGCAACAAGGACTGGCGAGGAGGTCGTGGGGCTCAACAAAACATCATTCCCT //660
  G  P  S  N  K  D  W  R  G  G  R  G  A  Q  Q  N  I  I  P  S
CATCTACAGGGGCTGCTAAAGCTGTGGGGAAAGTCATCCCTGCTGTTAACAACAAACTCA //720
  S  T  G  A  A  K  A  V  G  K  V  I  P  A  V  N  N  K  L  T
CAGGCATGGCTTTTAGAGTTCCTGTTCCTGATGTCTCAGTTGTGGATCTAACTGTCAGGT //780
  G  M  A  F  R  V  P  V  P  D  V  S  V  V  D  L  T  V  R  L
TGGAGAAAGGGGCAACTTATGATGAGATTAAAAAGGCAATTAAGGATGCCTCTGAGGGAT //840
  E  K  G  A  T  Y  D  E  I  K  K  A  I  K  D  A  S  E  G  Y
ATTTGAGAGGAATTCTGGGTTACACTGAAGAGGATGTTGTGTCTCAGGACTTCCTTGGAG //900
  L  R  G  I  L  G  Y  T  E  E  D  V  V  S  Q  D  F  L  G  D
ACCCCCGCAGCTCTATTTTTGATGCTAATGCTGGCATTGCCCTGAATAACAATTTTGTCA //960
  P  R  S  S  I  F  D  A  N  A  G  I  A  L  N  N  N  F  V  K
AGCTTGTGTCCTGGTATGACAATGAATATGGATACAGCAACCGGGTGGTTGAGCTCCTCC //1020
  L  V  S  W  Y  D  N  E  Y  G  Y  S  N  R  V  V  E  L  L  Q
AACATATGTATAAAGTGGATCACCAGTAGAGTTGAATCCAGCCAGTGAAGAATTTCAGAA //1080
  H  M  Y  K  V  D  H  Q  .
ATCTGATTGTGTAACTGCAAGATGTTGGAGTTTGGTGGATCCTTGCTGTGCTTATAAGTA //1140
CCCAAGTATTTGCAGACATCTTTATGGGTCATGATAATGAGTGTTATATGTATGTGAAAT //1200
GGAGTCAGATAATCATTTGTGTGTCAGTGGAACTTCCCATCAGTATTTCCCCAAATTTAT //1260
AGGCACACAACCAATTTTTTCACTGCAATTTTAAGCTTGCAAGATATTACTGGATAAAGA //1320
TTTACAATTTACTTGTTGGACCAAAGAACGTAGCTTGGCACCCATCTAAGAAAGGTTATC //1380
GACACAAGGGATGTTATCAAGACTGAAAGGGTTATTTCCTAAACAAACTTGCAGTTGTTC //1440
AAGTTAGCAGAATATAGCTAGCTCATTTTCAATTTATTTCTGAAAATGTCAACCCGTATG //1500
TATTTGGAGCCACTGACCTACGATTGACTCATCTGCTGTCCTATCACAGACCTCTAATGT //1560
ATTTGTTCACTTCCTTTAATTAGTTCCAGCTTTAGGTCTAGAACTATACAACACTTATAC //1620
ATTTGTGCTAAATGTATCTTTGCTGTCATGTGGGTTTTATTTGGCTTCAGTCCCTGTCTA //1680
GATGTGTCATGTGGGTTTTATTTGGCTTCAGTCCCTGTCTAGATGACAAATAATGGTGTT //1740
GGTAATTTAGATGAACTTTAATTTCTGTTTCAGGCTTGTACCCTAATACCAGCTAGTCGC //1800
TTTTAAAATAAAATGTGAAACCTTCAATGTTCTCATTTTCAGGTGCTTTTATATTAAAGT //1860
GCTGACCTTTCCCACCAAGTTGCTTGGGTTGTGAAGAGTGACATTAATAAAATAAAAGTG //1920
GCCAATGTTTATTTGACTCCAATACACTTAAAGAACGTGTTTGTATATTTTTCTAGTTTA //1980
CTTTTTTAAAAAGGAGATATTACATTGAGTTTTGTTGCACTAATTTGTTTAGTCGTCAGC //2040
TATCGCTTTAGATTTCAGCATTTCTATTTTTTAGACATGACGAGACTTGACGATCCACCA //2100
ATGAATGCAACAGATGTATTTTTAAATCTGGTCCATACATTAATTCTTTCTAGCATGGAC //2160
AAGGTAGTCAGGTGTTGAATTAACTTTTGTTTTAATTCAATTTGGGGGAAATGCTGACCA //2220
ATACCTTTTCTACCACACTATTGGTTAAGCATTTCAACTCTATAATTTGTTTAGTTTGGC //2280
CAGTACTATCCTCATATGACAGGTTACCATCTTAGCTTATGGAATGGATTGGATAAAGGT //2340
CTCGTCTCCAACATAGCTCATTGCACCTTTTGAAAAGCACATAGTTCCATCTAATTAAGT //2400
GGCATGTCTGTAGATATACATTGCCCAAGATCACCCTCTTTATGAAATAATAAAATACAA //2460
```

GAPDH RNAi Fragment

5' Primer Region

3' Primer Region

Figure 52 : GAPDH RNAi fragments

The figure shows the cDNA sequence and predicted translated region for a cathepsin L homologue previously identified. 5' and 3' Cathepsin fragments for insertion into RNAi vector are highlighted in blue and orange respectively. Regions primed by 5' and 3' primers for each fragment's initial amplification from cDNA are highlighted in green and red respectively.

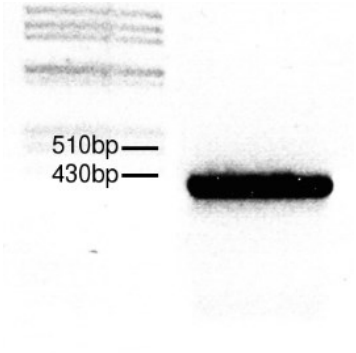


Figure 53 : GAPDH PCR product for insertion into RNAi transcription vector

GAPDH PCR product was amplified from first strand digestive gland cDNA and is predicted to be 373bp in length. The resulting fragment of amplification is shown here run on 0.8% Agarose gel

not have worked simply because of the lack of apoptosis machinery available in an adult *D. reticulatum*, due to the organism needing no further developmental changes. A potential future study could be the use of APIN RNAi against *D. reticulatum* eggs, but serves less application for molluscicidal agents, with eggs being a more difficult target to administer in the field.

5.4 RNAi against GAPDH

Having failed to show a phenotypic effect with RNAi of apoptosis inhibitor, after failing to demonstrate consistent down-regulation with RNAi directed against Cathepsin L, experiments to show that an RNAi effect could be elicited in *D.*

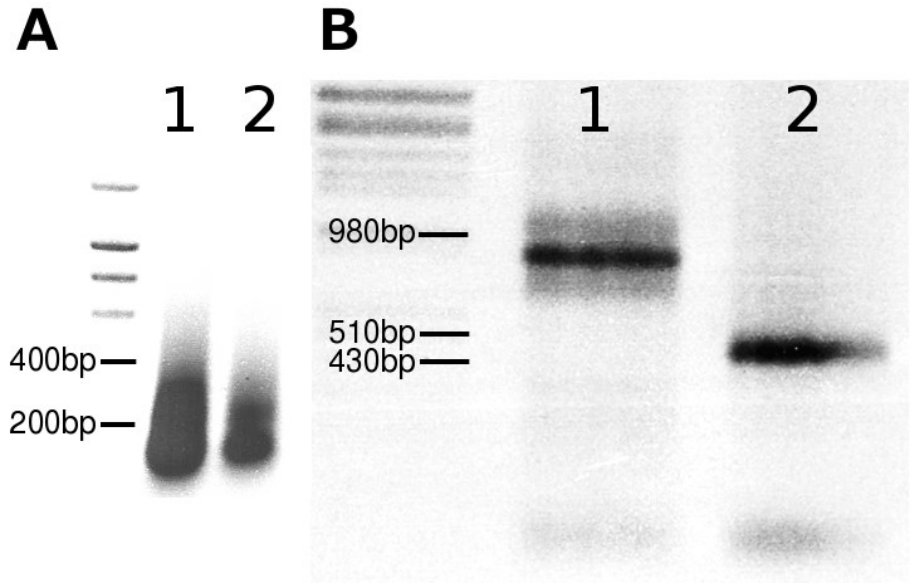


Figure 54 : ssRNA and dsRNA synthesised through in-vitro transcription of GAPDH RNAi Construct

A) Lane 1 and 2 contain 5' and 3' GAPDH construct ssRNA, the banding appears between around 200-400bp as expected. ssRNA often appears smeared on non-denaturing gel. Size was confirmed by assessment of dsRNA produced by annealing the two ssRNA samples together, gel B. Kanamycin ssRNA can be seen in Figure 45

B) Lane 1 contains the Kanamycin dsRNA synthesised alongside Lane 2 which contains the GAPDH fragment dsRNA. Both bands appear clearly at expected relative positions with limited low molecular weight degradation. The bands run to slightly different positions than the PCR equivalents due to dsRNA having a different charge to the DNA molecular weight ladder used here.

reticulatum were undertaken. Using a HKG as a target had the benefit of reducing the variability in gene expression from organism to organism seen with Cathepsin L. HKG are often used in RNAi studies as a positive controls with commercial technologies such as Life Technologies Silencer® RNAi positive control kits using GAPDH and 18S rRNA genes for human cell lines. EF1-A was used as the HKG, as it showed to be the least variable for whole tissue; additionally geNorm which used a

pairwise comparison method showed GAPDH vs. EF1-A to be the best combination for demonstrating changes in expression.

A fragment of *GAPDH* gene was amplified with PCR from first strand digestive gland cDNA previously synthesised, section shown in see Figure 52. The predicted PCR product of 373bp was size confirmed with gel electrophoresis, see Figure 53. The product was gel extracted and restriction digested and ligated into the pLitmus28i vector. This was then used in for *in-vitro* transcription reaction with the resulting ssRNA and then dsRNA purified and confirmed with gel electrophoresis, see Figure 54, then used for RNAi assays with *D. reticulatum.*

5.4.1 *GAPDH* dsRNA Injections at 20µg

Injections of 20µg of GAPDH dsRNA and Kana dsRNA (at 4µg/µl), along with Sham (Buffer Only) injections were conducted on 15 slugs. They were observed for phenotype change for 6 hours, then after 24 hours flash frozen in liquid nitrogen. Total RNA was extracted from whole organisms. First strand cDNA was produced and subsequently used for qPCR analysis to assess for gene knock down. The resulting data shown below indicated no change in GAPDH gene expression as a result of dsRNA injection (Figure 55). Since no indication of an RNAi effect was observed, further time points were not taken, but instead altering of the experimental conditions was considered.

5.4.2 GAPDH dsRNA Injections at 50µg with *in-vivo* produced dsRNA

In order to test whether larger quantities of dsRNA would produce a gene knock-down effect dsRNA was produced *in vivo* via a method described in the literature (Solis et al. 2009). The use of *in vivo* methods allowed much larger amounts of dsRNA to be produced. The GAPDH fragment and Kanamycin gene previously in Litmus28i were cloned into the L4440 vector and transformed into a RNaseIII deficient *E. coli* strain HT115 (rnc14::ΔTn10) which expresses T7 RNA polymerase under IPTG induction. dsRNA was purified from bacterial cultures after induction, with some optimisation for GAPDH (see Methods section, Ch. 2), see Figure 56. The resulting dsRNA was checked via gel electrophoresis to confirm size

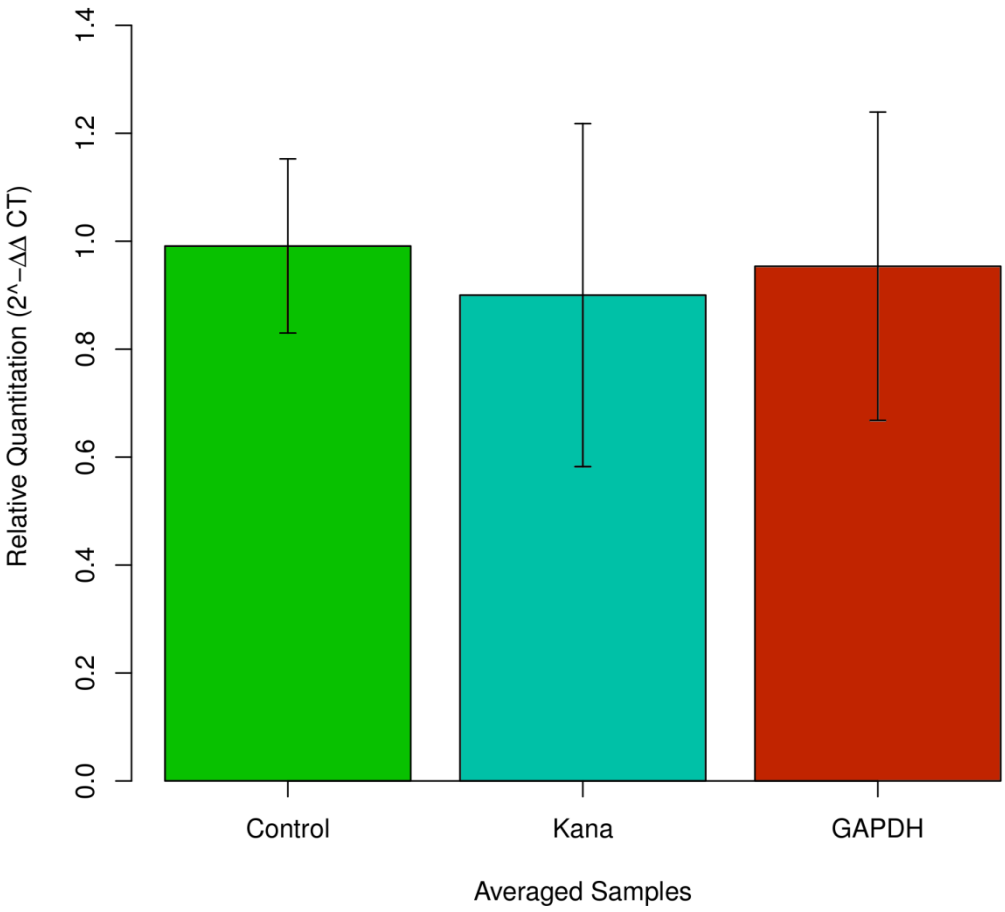


Figure 55 : Relative quantitation for GAPDH RNAi study at 20µg

Figure shows 3 averaged samples, measuring GAPDH levels vs. EF1-A control gene. Control: average of 4 sham injected individuals. Kana: average of 5 individuals injected with Kanamycin gene dsRNA. GAPDH: average of 5 individuals injected with GAPDH fragment dsRNA. No noticeable effect was seen across any of the samples although internal variation was relatively high as shown by the error bars which represent 1 standard deviation of the biological replicates used to generate average values. Independent T-Test indicated no significant difference, with p -values for Control vs GAPDH and Kanamycin vs GAPDH at 0.82 and 0.78, and Control vs Kanamycin at 0.62.

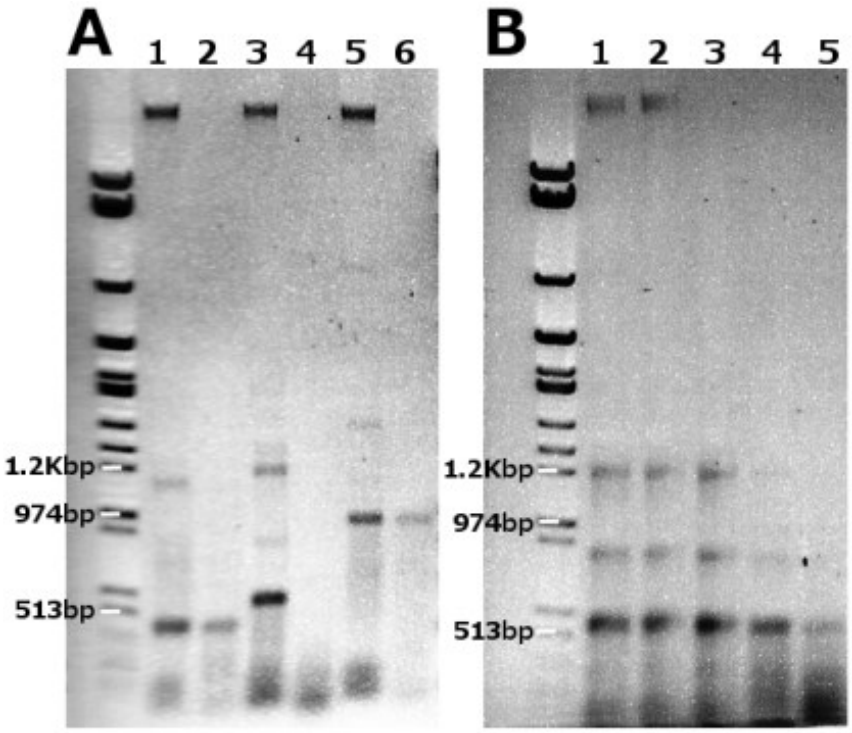


Figure 56 : GAPDH in vivo dsRNA production

A) Lanes show nucleic acids extracted from the in vivo expression system transformed with RNAi constructs before and after nuclease treatment and purification. Lane 1 shows APIN before nuclease treatment and lane 2 after treatment. Lanes 3 & 4 and lanes show GAPDH and 5 & 6 Kanamycin before and after nuclease treatment. Notably lane 4, GAPDH after nuclease treatment has no observable band at the expected region. Nuclease treatment was further optimised for the GAPDH construct.*

B) Lane 1 shows GAPDH nucleic acids before nuclease treatment and lane 2 contains control, nuclease treatment with no nuclease. DNase was previously tested and concentrations did not affect visibility of expected dsRNA band. Lanes 3, 4 and 5 represent 10%, 25% and 50% of the original RNase A treatment. Lane 4, 25% RNase was considered the most optimal treatment and GAPDH dsRNA was purified using this RNase A concentration.

**in vivo synthesised APIN dsRNA was not used, only in vitro dsRNA was used for APIN experiments.*

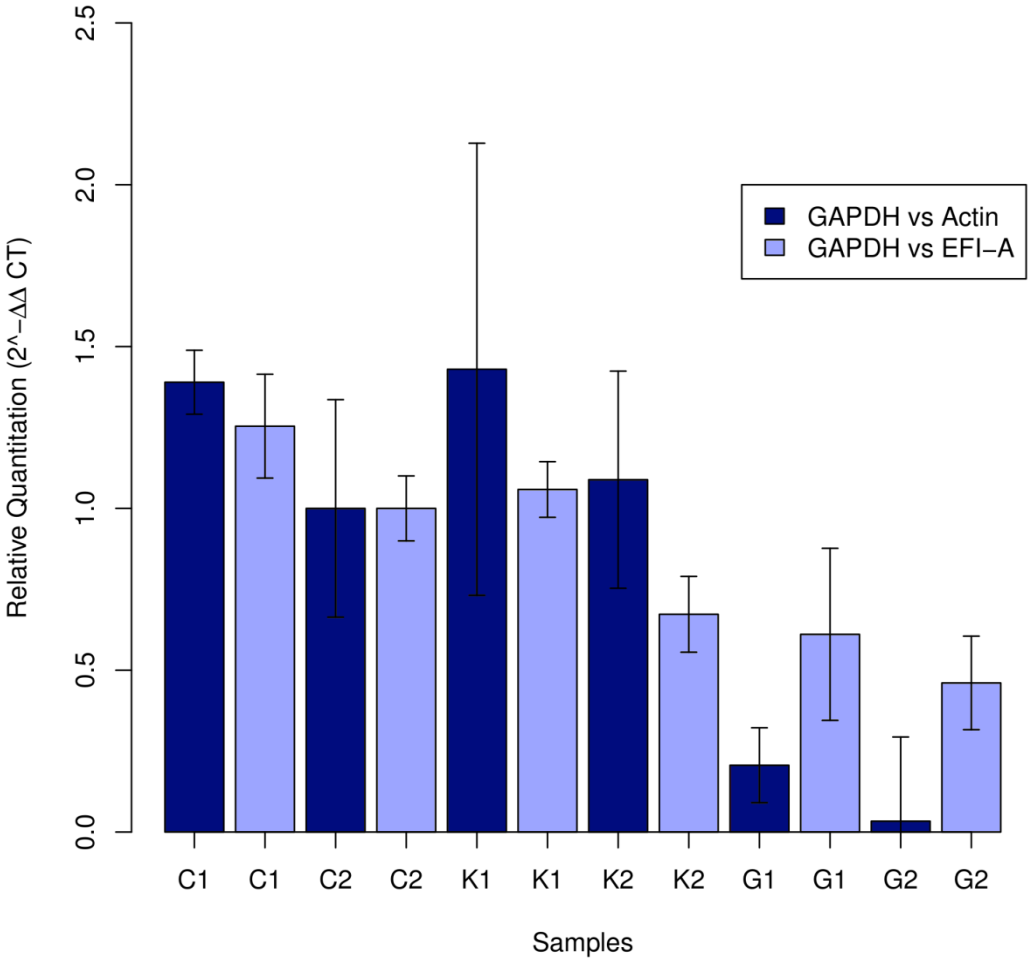


Figure 57 : Relative quantitation of GAPDH vs. Actin and EF1-A

Each sample represents a pool of 4 slugs injected with either buffer (C1, C2) 50ug kanamycin dsRNA for each individual in the pool (K1, K2) or 50ug GAPDH dsRNA for each individual in the pool (G1, G2). The two colours represent the different quantitative PCR studies for each sample, GAPDH with Actin HKG primers (dark grey) or EF1-A HKG primers (light grey). In general comparison with Actin show favourable results, with decreases in GAPDH to <20%. However comparison with EF1-A, whilst correlates to some degree, shows much greater uncertainty. Error bars represent 1 standard deviation of the 3 technical replicates for each sample.

and appeared equivalent to *in vitro* transcribed dsRNA shown in Figure 54. The resulting GAPDH and Kanamycin dsRNA was injected at 50µg doses (5µg/µl) into 8 individuals each, along with 8 individuals sham injected as a control. The individuals were monitored for 7 days for any observable phenotype occurring. At the end of 7 days 8 slugs from each injection group were split into 2 pools and RNA extraction of total of 6 samples was conducted, each sample representing 4 individuals. The resulting RNA samples were used for quantitative PCR. Figure 57 shows the resulting data normalised both with EF1-A and Actin based on Ct values shown in Table 26.

When the GAPDH expression data are normalised to Actin HKG there is convincing evidence of gene knockdown of GAPDH, with both RNA samples showing <20% of the lowest control expression. However when expression data are normalised against EF1-A much greater uncertainty with regard to any gene knock-down effect is apparent. Whilst the results of both genes show correlation, see Figure 58, the correlation is not strong and suggests one or both HKG's have variable expression. T-tests of replicates indicate no significant differences (p-value < 0.05), whilst t-tests of RNAi variables indicate significant differences between both control and kanamycin control when compared to GAPDH, but only with Actin as a HKG. When using EF1-A as HKG, the difference between GAPDH and kanamycin control is no longer significant, see Table 27-Table 29.

These results appear promising, with a consistent trend to down regulation in GAPDH expression after injection of GAPDH dsRNA. However, in view of remaining variability in the data, the conclusion that should be drawn is that, until better validation of a reliable HKG for each tissue in use can be demonstrated, qPCR will not be a reliable method for proving an RNA interference effect is occurring in *D. reticulatum*. RNAi components such as Dicer RNase III and Argonaute are not present in the available *D. reticulatum* datasets. However homologues of both these genes in the *C. gigas* and *A. californica* sequence data (GIs: 524901378, 405970135, 405960420, 524899606) are present and indicate the Molluscan phylum should be susceptible to RNAi. RdRP can be found for in *C. gigas* (GI: 405952885), which indicates there may be a systemic RNAi effect as *Mollusca*.

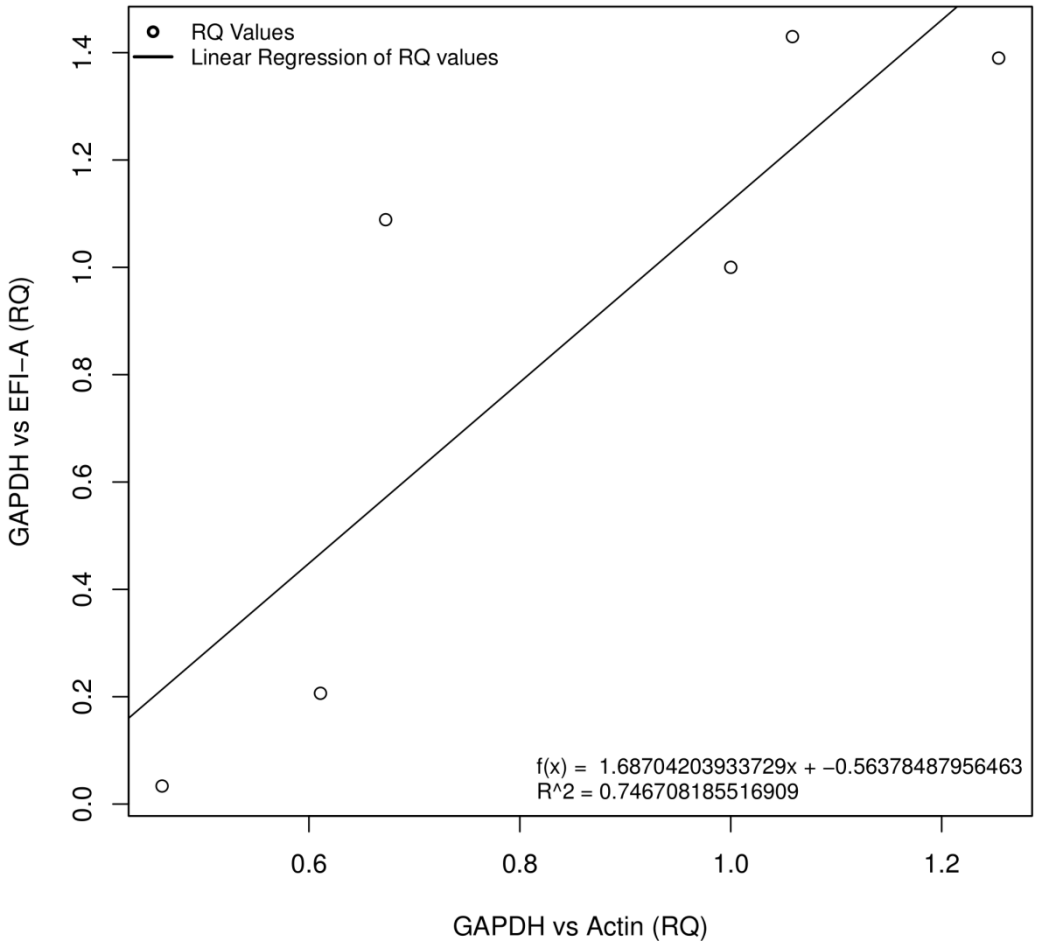


Figure 58: Linear Regression of GAPDH VS HKG

R^2 valued indicates while the normalisation of GAPDH with each housekeeping gene does positively correlate to a lesser degree. The relationship is poorer than expected and indicates one or both HKG may not be reliable enough to draw conclusion from.

Sample	**GAPDH**	**EF1-A**
Control 1	22.1	19.7
Control 2	21.9	19.1
Kanamycin1	22.4	19.7
Kanamycin2	22.8	19.5
GAPDH1	23.0	19.5
GAPDH2	23.8	19.9
Sample	**GAPDH**	**Actin**
Control 1	28.5	20.7
Control 2	27.4	19.2
Kanamycin1	27.1	19.9
Kanamycin2	27.7	19.6
GAPDH1	30.1	19.6
GAPDH2	28.6	20.4

Table 26: Ct values for two qPCR studies on RNAi data

Ct value for qPCR data showing comparative Ct between GAPDH and EF1-A and GAPDH and Actin. Initial qPCR runs were used to identify optimal starting cDNA quantities for reactions so that Ct values, wherever possible, fit into the 20 < Ct <30 range, which is considered most accurate. In the case of Actin starting cDNA was 5ng per well (compared with EF1-A at 40ng).

Compare Biological Replicates		Independent T-Test
Actin		
C1 vs. C2	1.39	1
G1 vs. G2	0.21	0.03
K1 vs. K2	1.43	1.09
	Result (p):	**0.60**
EFI-A		
C1 vs. C2	1.25	1
G1 vs. G2	0.61	0.46
K1 vs. K2	1.06	0.67
	Result (p):	**0.35**

Table 27: Comparison of Biological Replicates for GAPDH RNAi

Independent T-Test was conducted to assess whether there was a significant difference between the biological replicates where significant is considered to be $p < 0.05$. In this case biological replicates for both HKG are not a significantly different between biological replicates p-values 0.6 and 0.35.

Compare HKG Replicates		Paired T-Test
	Actin	EFI-A
C1	1.39	1.25
C2	1	1
K1	1.43	1.06
K2	1.09	0.67
G1	0.21	0.61
G2	0.03	0.46
	Results (p)	**0.92**

Table 28: Comparison of HKG Replicates for GAPDH RNAi

Paired T-Test was used to assess whether there was a significant difference between the 2 housekeeping genes where significant is considered to $p < 0.05$. In this case there is no significant difference between housekeeping genes used, p-value 0.92.

Comparison of Variables				**Independent T-Test**
	Control	GAPDH		
Actin	1.39	0.21		
	1	0.03	C v G (actin)	0.04
EFI-A	1.25	0.61		
	1	0.46	C v G (EFI-a)	0.06
			CvG Overall	**0.002**
	Control	Kana		
Actin	1.39	1.43		
	1	1.09	C v K (actin)	0.83
EFI-A	1.25	1.06		
	1	0.67	C v K (EFI-a)	0.38
			CvK Overall	**0.6**
	Kana	GAPDH		
Actin	1.43	0.21		
	1.09	0.03	K v G (actin)	0.03
EFI-A	1.06	0.61		
	0.67	0.46	K v G (EF1-a)	0.25*
			KvG Overall	**0.01**

Table 29 : Independent T-Test of Variables for GAPDH RNAi

*The table shows an independent T-Test to ascertain whether the groups Control, Kanamycin and GAPDH are significantly different. With each comparison CvG, CvK & KvG a T-Test for each HKG is conducted, then an overall including data from both HKG. For overall T-Tests the data suggests a significant difference between the GAPDH RNAi injected group and the Control group and the GAPDH group and the Kanamycin control group p-values 0002 & 0.01 respectively. And there is no significant difference between the Kanamycin and the Control, p-value 0.6. *However when we look at individual primers we see that K v G using EFI-A does not follow this trend, and the test indicates the difference between Kanamycin and GAPDH is not significant when using EFI-A as the HKG p-value 0.25.*

5.5 TNF Ligand-like gene as a molluscicidal agent

In light of a lack of sufficiently promising results with RNA interference, alternative mechanisms for causing lethality, the end goal of developing novel molluscicides, were considered. The *D. reticulatum* contig C2550;S1758;N554d showed similarity to a number of Tumor Necrosis Factor ligands (TNF-L). Unlike the C1q proteins which can also bind to cell receptors to elicit apoptosis, TNF ligands are usually homotrimeric requiring only a single peptide to form the complete complex. As cytokines, they have a range of potential functions, but can cause cytolysis or cell death. A previous study with the disk abalone *Haliotis discus discus* showed expression of a Fas-like ligand, isolated from cDNA, in an *E. coli* expression system. The study went on the show phenotypic effects in the target mollusc (De Zoysa et al. 2009). Whilst not proving the ligand is specific to mollusc it does present a potential opportunity. TNF-like ligands in *D. reticulatum* could be expressed and assayed against the organism for toxicity. If the protein can be shown to be target specific it provides a potential avenue for a molluscicidal product. Using the organism's own protein may also have the added benefit of decreased resistance with resistance leading to the resistance to its own cell signaling molecules. As both an exemplar of how sequence data could identify targets and as a potential opportunity to produce a molluscicidal agent, the TNF like sequence was expressed as a recombinant protein.

5.5.1 Recombinant Protein Expression

The TNF-like ligand from *D. reticulatum* (dTNF) was expressed as a Thioredoxin (TRX) fusion protein using the PET32a expression vector in *E. coli* Origami B(DE3) cells. Thioredoxin are ubiquitous proteins, 11.9kD in size, containing catalytically active disulphide group and function in several biochemical pathways by thiol/disulphide exchange reactions (Lemaire et al. 2000). Thioredoxins are structurally stable, and remain stable in *E. coli* even under extreme conditions (Holmgren 1985) as well as promoting protein folding of fused proteins during expression. The full coding sequence of dTNF minus the transmembrane region was inserted in-frame after the TRX site (see Figure 59). The transmembrane region was

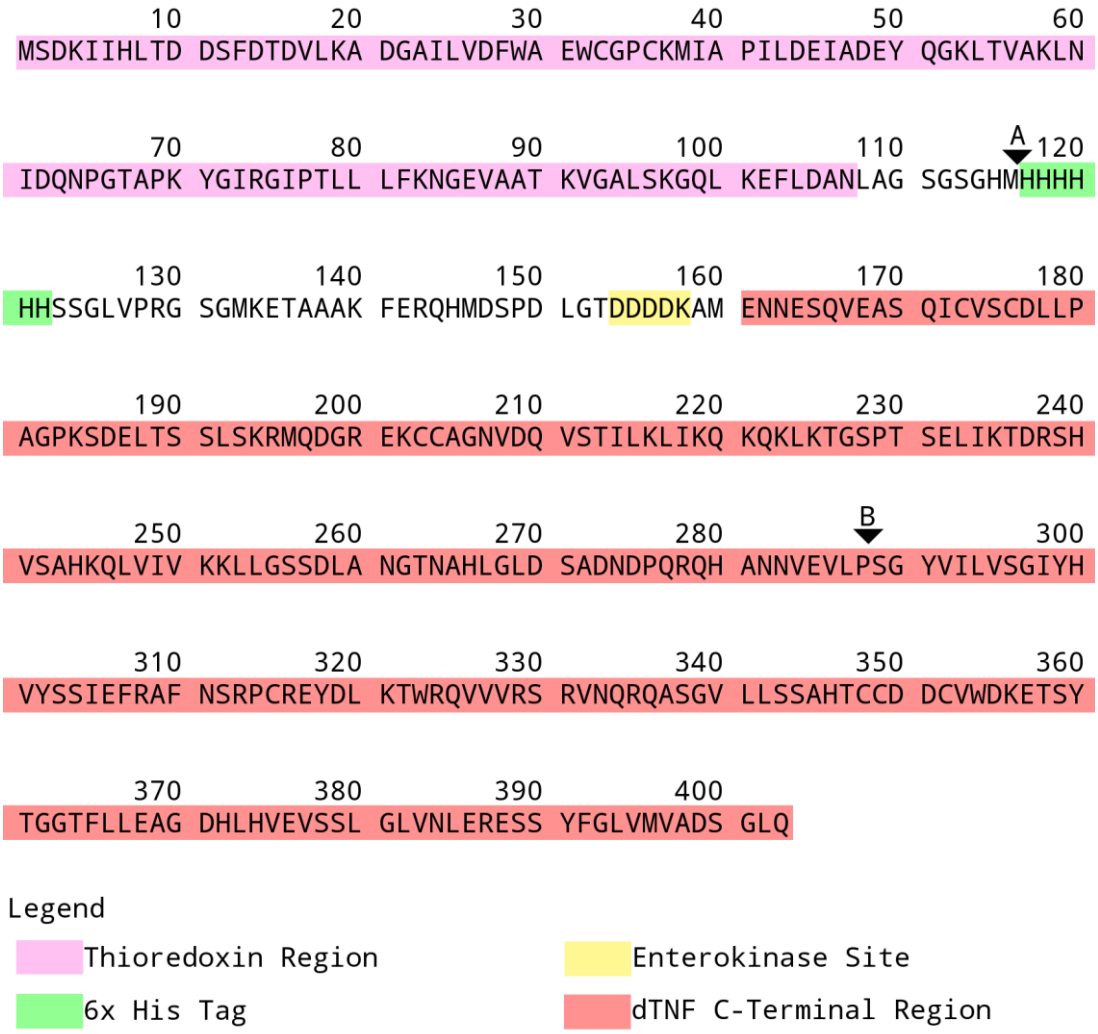


Figure 59 : Protein sequence of Trx-dTNF with highlighted regions

Trx-dTNF protein sequence, the complete protein is predicted to be 43.9Kda. Site markers A and B are positions where cleavage at those points would produce a ~31Kda peptide. The sizes of highlighted regions are: Thioredoxin 11.9Kda, His Region 5.3Kda, and TNF 26.7 KDa.

removed to avoid interaction of the protein with the host expression system's own membrane and improve protein solubility. The PET32a vector contains 2 potential 6xHis tags, the first 3' of the TRX site, 5' of the MCS, the second is 3' of the MCS this was not included in the final protein via introduction of a stop codon, in order to avoid blocking dTNF's potential receptor binding function which is at the C-terminal

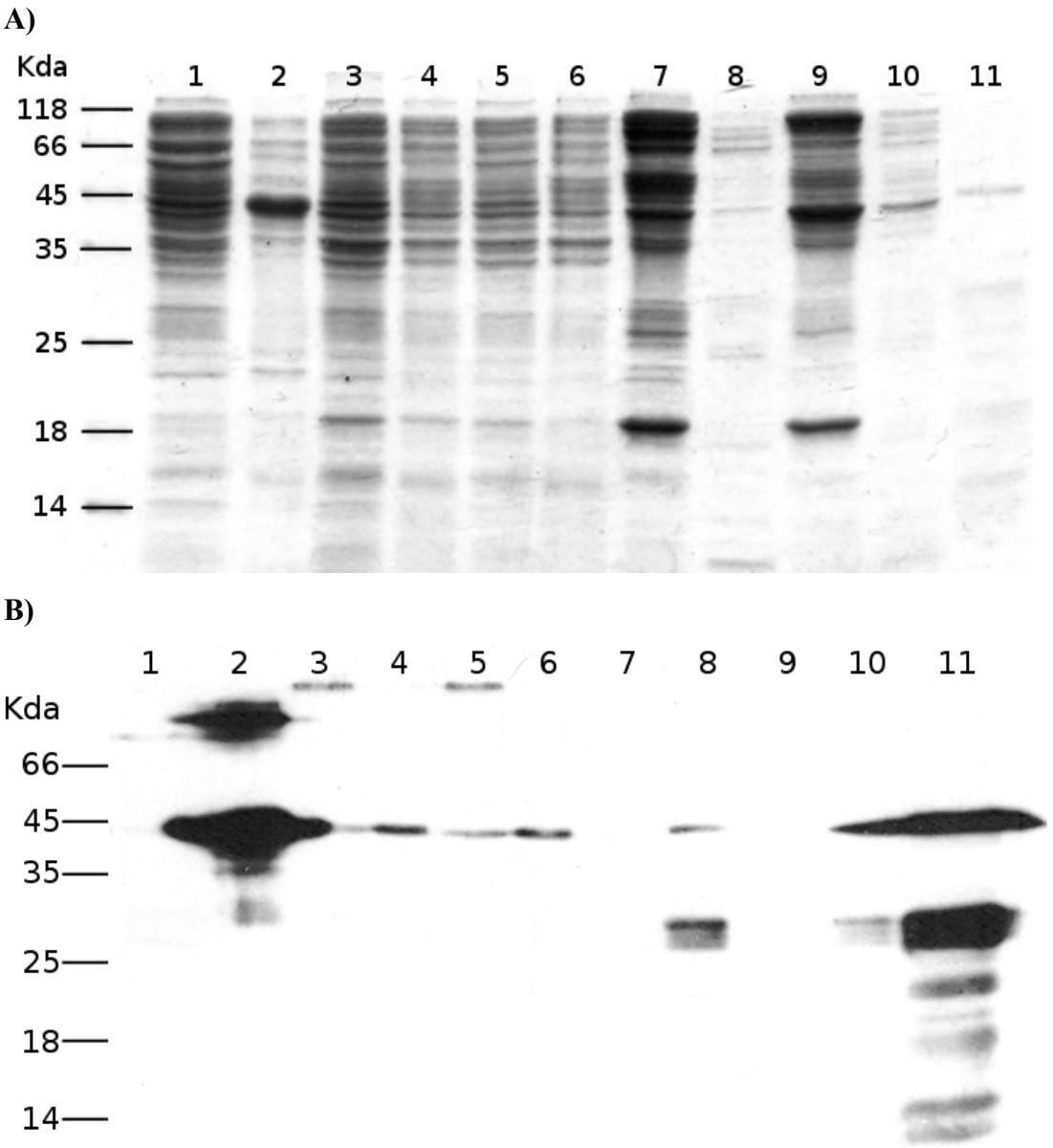


Figure 60 : CBB and Western blots showing purification steps for TRX-dTNF

Figures A and B show the two gels loading with equal amounts of protein and run in parallel, A was stained and B was blotted with Anti-His antibodies. The Western blot indicates an immuno-reactive protein at around ~45Kda in all lanes apart from 7 and 9, which are the column wash elutions. A large amount of 45Kda protein is seen in the uninduced insoluble fraction, lane 2. A smaller amount is seen in the uninduced (lane 1) and induced soluble lane (3, and 5) and the induced insoluble fractions (4 and 6). Lanes 8 and 10 are the column elution fractions and lane 11 shows elution after dialysis and freeze drying

end of the peptide. The aim of this construct was to minimise the potential side-effects of fusing dTNF, by replacing the transmembrane with a region of similar size (TRX) and avoiding interfering with the receptor binding end. The calculated molecular weight of the Thioredoxin dTNF fusion protein (Trx-dTNF) was 43.9 KDa.

After transformation of *E. coli* origami B(DE3) cells [Novagen] with the expression construct, transformed clones were selected. Cell lysates were analysed for the presence of TNF, using immuno-reactivity to Anti-(His) antibodies as a marker for the expressed protein, Figure 60. Expression of recombinant TNF was found to be higher in cultures grown without induction with IPTG (see Figure 60, compare lanes 1 & 2 uninduced soluble and insoluble versus lanes 3 & 4 soluble and insoluble); in the insoluble fraction from uninduced cells, TNF was present as a major band of approximately 44kDa. The reason for this expression in the absence of induction is not clear, though there is a cessation in culture growth after the introduction of IPTG not seen in uninduced cultures. No overall change in OD after IPTG induction, compared with increase in OD seen in uninduced cultures, suggest that the induction process may have had a toxic effect on the bacterial culture. Alternatively in some cases uninduced cultures can auto-induce after the supply of none-lactose sugars runs out, expressing proteins several times higher than induced equivalents (Studier 2005). Several expression cycles were conducted with similar results. The $(His)_6$-tagged protein was purified by metal affinity chromatography using HisTrap HP column. As seen in lanes 8, 10 and 11 this resulted in some degradation of the recombinant protein; bands are seen around 30 KDa, which are immuno-reactive with the anti-(His) antibodies, and can only be either Trx-His-Partial dTNF or His-dTNF, see Figure 59 points A and B. Elution salts including imidazole were removed from the protein purified by affinity chromatography by either dialysis or buffer exchange columns; in both cases precipitate formed during the purification step, with resulting soluble and insoluble fractions both containing Trx-dTNF. The final protein used in assays was u used the buffer exchange method as this gave better yields, and the buffer could be used as a control in subsequent experiments. Additionally, the buffer exchanged purified protein appeared to be far less degraded, with almost all

immuno-reactive protein being around ~45Kda and only a small band at around 30Kda. However there was still a significant amount of background none immuno-reactive protein in the purified sample, Figure 61. The final purified protein was resuspended to 5µg/µl which was quantified based on the CBB gel, with the actual total protein concentration being 33µg/µl.

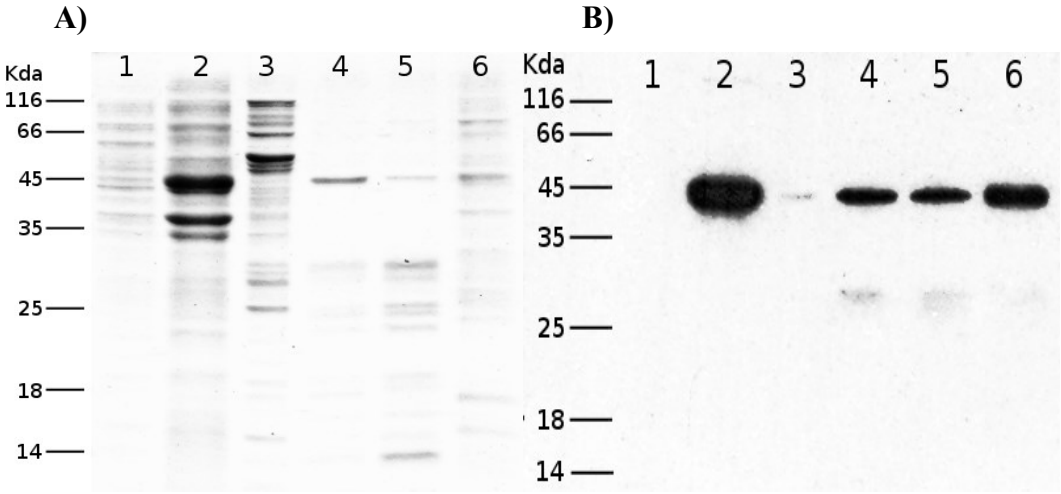


Figure 61 : Coomassie brilliant blue and Western blots of buffer exchange purified dTNF

Blots A and B represent the CBB and Western blots respectively. Lanes 1 and 2 compare induced and uninduced fractions, we see no evidence of any protein in the induced fraction, even more so than previous induced. Lane 3 is the His-Trap column wash and 4 and 5 2 elutions from separate loadings of the soluble fraction of the uninduced culture sonication. Lane 6 shows the final protein product after purification with almost no immuno-reactive protein outside of the ~45Kda band, suggesting limited degradation of dTNF. Small amounts of immuno-reactive protein is seen around the 30Kda mark, mainly in lanes 4-5, as seen in previous blots but to a lesser extent.

5.5.2 Injection Assay

10 slugs were each injected with 50µg (330µg of Total Protein) of Trx-dTNF in 10µl of TAE buffer, buffered to 7.4. As a control injection, 5 slugs were injected with imidazole elution buffer, purified in the same manner as Trx-dTNF was

conducted in parallel. Previous injections (data not shown) of 500μg of Ovalbumin indicated that individuals should be tolerant to injections of high levels of protein. Individuals were monitored for 2 weeks after injections and no phenotype was observed. The lack of results and availability of time left in my project, lead me to stop investigation at this point. There are a number of further pieces of work that could be done with regard to Trx-dTNF; the possible reasons for lack of activity are numerous. However one of the main reasons for producing the protein was as a prototype to show the potential of expressing native *D. reticulatum* proteins which could be tested for toxicity.

Chapter 6 | Discussion

6.1 Assembly Comparisons

The number of contigs with significant similarity to database sequences by comparisons such as BLAST is commonly used as a metric summarising transcriptome data and indicating how much of the dataset is likely to represent actual transcripts. The number of contigs and number of unique protein matches shown in tables 13 and 19 are highly varied from assembly to assembly. Using this metric as the primary arbitrating factor over the homology summary would change the results by 2 to 3 times depending on which assembler you were to use. However the overall number of reads with BLAST matches is much less varied and top IPR terms seem also to be very similar for most sequences of the digestive gland tissue. When we consider the overlap of reads between assemblies which have BLAST matches (Figure 20) we see that majority ~80% are the same. This seems to indicate the majority of the dataset represents the same group of sequences indicated by the IPR terms, regardless of how the data is assembled. The minority of the dataset is split over a disproportionately large number of contigs which varies greatly between the assemblies and skews any statistics tied to contig numbers rather than read numbers. However for neural tissue contigs the parity between assembly homology data is poorer reflecting the likely decreased quality of the source data, and in turn the cDNA used for sequencing. In conclusion the assemblies, whilst important for ascertaining complete sequences of genes, can significantly affect the outcome of analysis when considered on their own. Considering multiple assemblies in parallel with using read numbers rather than contig numbers arguably biases the analysis.

In addition to overall analysis, utilisation of multiple assemblies produces overall the largest number of genes available to researchers. We find in some cases 1 or more assemblies do not contain a gene which, when examined manually by simply looking at the sequence and the homology matches is obviously a real gene. Examples of this failure of assembly software include several C-type lectins, which on investigation by hand are evidently unique genes, despite not being present in some of the assemblies. In other cases such as ferritin gene, one assembler may have

incompletely assembled the gene, but by comparing the internal homologues between assemblies we can get the most complete gene, which can be confirmed by comparison with external homologues and if necessary by hand. In previous studies assemblies have been further assembled with a second pass, merging multiple assemblies into one. This has the disadvantage of often further over-assembling contigs which may already be over-assembled. For the purposes of our analysis there was no necessity to merge assemblies as they could be interacted with as a 'meta-assembly' in a BioSQL database. However most external databases would only accept a single assembly file for publication. In this case the temptation would be to 'cherry-pick' contigs from our current dataset. However such a workflow, despite being relatively easy to setup, would require its own assessment and evidence of improvement over a single assembly. This is beyond the scope of the current project. As such contigs from the CLCbio assembly were uploaded to the NCBI database.

6.2 Use of contigs, reads and coverage

A major advantage of analysing high-throughput sequence data is the ability to predict things like gene function. Here we attempt to make further predictions using read numbers to weight the significance of predicted genes when considering an overview of the transcriptome. Correlating sequencing frequency with expression levels has been described in a number of papers, increasingly replacing qPCR/microarray methods (Sultan et al. 2008; Schmidt, Schmid, and Grossniklaus 2012). However read numbers or reads per kilobase of exon model (rpkm) values usually rely on a genome to map transcripts to genes to bypass the problems of incorrect assemblies.

We recognise that frequency of a transcript being sequenced does not necessarily indicate that the transcript exists in larger quantities than others, but argue that it is a good indication. Also, as previously described, examples such as digestive enzymes homologues cathepsin L and cellulase may have more transcripts but could conceivably have less biochemical significance due to poorer enzymatic efficiency or greater post-transcriptional regulation, rather than having a more significant role in the digestive action of the organism. However as not all sequences

can be considered within this body of work, when overviewing sequences prioritisation is necessary. Linking reads to IPR terms has been done in previous studies (Peng et al. 2011) and both avoids prescribing expression to genes whilst still providing some sort of weighting to indicate prevalence. Previous methods of ranking top IPR terms based on the number of contigs linked to IPR terms such as described in Pauchet et al. (2009) seem likely to bias towards larger gene families and sequences which are fragmented into multiple contigs.

6.3 Homology Analysis Caveats

The use of homology allows for prediction of protein function on a scale much larger than can be achieved using an experimental characterisation. However inevitably the function is only a prediction and relies upon probabilities, with the guarantee that certain percentage of predicted protein functions are incorrect. A problem with BLAST-based homology is the inclusion of predicted proteins within the NCBI databases. This causes an observable cascade of predictions strengthening further predictions. This is particularly problematic for phyla such as Mollusca where the majority of close homologues are from equivalent mass sequencing efforts. In this case, examples where a *D. reticulatum* protein had homology to a series of different accessions, all of which were predicted proteins based either on each other or eventually an experimental derived protein often from distantly related model organisms, were numerous. The actual homology of the *D. reticulatum* protein to any protein which was not a predicted protein, but had an experimentally derived function was often much lower than to the most similar protein with no experimentally derived function.

In an attempt to reduce this effect, our discussion of predicted protein functions was primarily based on INTERPRO data. Whilst this also has caveats, the homology is based on group patterns from a large pool of sequences rather than a single source-based prediction. The data is better curated, and a much higher percentage of the proteins have been experimentally derived. Despite this there is obviously still bias within the data, an example of which is the spider toxin function. Without human intervention an automatic assignment of function would label the *D.*

reticulatum peptide a spider toxin. It is much more likely that spider toxins are the only proteins which use the cysteine knot domain to have been well described and included in INTERPRO. Despite this it still gives clues as to what potential role the protein has, as discussed in chapter 4, which is the primary goal of the homology analysis.

6.4 Programming Tools

Whilst the linked toolkit represents quite a large amount of programming code, it is not the focus of this thesis. The toolkit is primarily a collection of scripts for dealing with biological data and interacting with the BioSQL database. A large amount of the tools are simply automations of tasks that can be done by hand. In some cases they duplicate suitable functions which exist in main biology libraries such as bioJava and bioPerl, but were simply quicker to write than find, while in other cases, an alternative implementation was necessary. For example the bioJava BLAST parser is a 'push-parser' (whole file is 'pushed' into memory) rather than a 'pull-parser' (Pieces of file a pulled into memory); this method can be problematic for large XML files, and has been addressed for some specialist datatypes such as mass spectrometry (Griss et al. 2012). For large BLAST XML data files, the default parsing would fail to load due to bug 6536111 in the Java virtual machine [http://bugs.sun.com/bugdatabase/view_bug.do?bug_id=6536111]. In this case I used a 'push-parser' implementation were the file is loaded piece-meal into memory, using the Woodstox XML parser (‘Woodstox’ 2013). This had the added benefit of supporting BLAST XML files with multiple BLAST records, although there are several tools to split the file into individual records.

Other sections include bulk download scripts for all the sequences linked to a Panther, IPR, PFAM, NCBI, UNIPRO term, which were written to avoid additional steps. For instance a contig name could be used directly to download all relevant sequences from the matching IPR terms and automatically align them with clustalw with a single command. In many cases these tools represent workflows, for example taking a sequence, BLASTing it, retrieving related terms. The graphs showing BLAST cut-offs were done by counting the number of BLAST records then

decrementing that value for each e-value in order of e-value. In these examples it is difficult to decide to what extent to explain the automated functions as in many cases they would not need to be detailed were it done by hand. In my thesis I have neglected the explicitly explain what all the tools/scripts/functions/classes do and how they work. I have instead opted to make the source code available for all the tools used in my PhD as an open-source project at https://github.com/EnderDom/Eddie. This is a publicly accessible project, hosted by a popular source control website and allows for any part of the source code to be inspected and downloaded.

6.5 General Conclusion

The objectives of this work were to extract sequence information from a crop pest which previously had limited genomic data and explore the implications of that information. *Deroceras reticulatum* was studied to better understand the underlying molecular mechanisms with the long term goal of identification of effective strategies to combat the organism in its role as a crop pest.

In the initial stages of the project the aims were the production of RNA and then cDNA of sufficient quality to acquire genetic data from the species. The extraction of RNA differed significantly from previous work done with insects and techniques had to be adapted to better suit the Molluscan physiology. In general dissected tissues using fresh tissue homogenisation techniques produced overall better quality RNA than other methods. For experiments where RNA quality was paramount such as cDNA synthesis, dissected tissue was much preferred due to improved quality over whole tissue. This added further weight to initially focusing on digestive gland, alongside previous research on crop pests where digestive gland sequencing produced useful genetic information. Both RNA and cDNA protocols were repeated until the appearance on gel electrophoresis matched the expected appearance based on previous research and relevant manufacturers.

With the cost and general trepidation that comes with the use of new technology such as pyrosequencing, evidence that the cDNA contained *D. reticulatum* sequences was required. With almost no sequence information available,

using PCR to amplify known genes was not a feasible method. The production of cDNA libraries, the cloning of second strand cDNA into vectors and transforming into *E. coli*, was standard laboratory practice, in order to preserve any cDNA produced. Previous to high-throughput sequencing methods, cDNA libraries could be sequenced with Sanger techniques to provide transcript data. Sequencing a sample of clones containing cDNA was a clear method to check the presence of *D. reticulatum* sequences. After 12 clones were sequenced it was clear that the sequences belonged to *D. reticulatum*. BLAST homology of the clones showed sequences matching invertebrate and in many cases Molluscan homologues despite a general lack of Molluscan homologues in the NCBI database.

With the cDNA synthesis protocol optimised for *D. reticulatum*, investigation into ion channel proteins was carried out. Ion channel proteins were of interest in many crop species and isolating the sequence would have been of great benefit. The long term goal of the research group was to isolate ion channels from target crop pests which potentially bound to invertebrate specific spider toxins and conotoxins. These could then potentially be expressed in Xenopus oocytes, commonly done for ion transport studies (Wallingford et al., 2010), and used as an assay for the blocking of the ion channels by toxins. The result of degenerate PCR to amplify an ion channel transcript sequence was a partial fragment homologous to the sodium voltage gated ion channel protein of *A. californica*. RACE PCR was then used to try and recover the remaining sequence, but was unsuccessful.

It may be beneficial for future work to recover the remaining sequence information for the ion channel homologue fragment sequenced in this project. Degenerate primers could be designed to make use of the increased availability of Molluscan sequence data published since this initial experiment. The ion channel was not found in our neural dataset, possibly due low transcription levels which would be expected for ion channel proteins. cDNA normalisation might improve the overall likelihood of retrieving the sequence through high-throughput sequencing. Other alternatives include looking instead at genomic DNA and using a primer walking strategy for sequencing the rest of the gene (Leoni et al., 2008). Despite failing to generate the full transcript here, future work should be able to procure its

full length given enough time. The fragment generated by this project is a useful starting point and also gives an indication of the conservation of the gene in *D. reticulatum*.

With the cDNA validated, digestive gland high-throughput sequencing and analysis was conducted and confirmed much of the previous research of the organism's digestive biochemistry. We highlighted significant genes, where data was available, and gave weighted functional predictions of sequences found. The provision of major digestive enzymes in crop pests has been a first stage goal in previous invertebrate pests, particularly insects. For *D. reticulatum* this provision was done relatively swiftly in comparison to classical cDNA library cloning and Sanger sequencing, utilising modern high-throughput sequencing methods. With this major objective completed we considered further tissue types, which lead to sequencing neural tissue cDNA. With decreasing costs and improved sequencing technologies, other tissue types could be considered in the future. Identifying expression levels with sequencing data has already been described and could be used as replacement of micro-array analysis. Future research may be able to link sequencing data together to build an expression profile for *D. reticulatum* tissues, as has been done in model organisms.

RNAi against the Cathepsin L homologue was the first implementation of the sequence data analysis within this project. Previous studies had been done on insects using RNAi against cysteine proteases, such as in pea aphid (Jaubert-Possamai et al. 2007), and the general interest in inhibiting proteases of crop pests lent credence to the potential of the experiment. However a number of differences lead to RNAi in molluscs being less convenient than insects. The lack of previous work and the lack of a general standard housekeeping gene have been identified within this work. As well as population sizes available for research use, being significantly lower. This lead to the studies being scaled down with less confidence in the qPCR results than with an equivalent qPCR in insect populations. Cathepsin L variability appeared too high to produce accurate qPCR results, without a phenotype the gene was not investigated further. Some improvements could be made such as having larger samples sizes for qPCR studies. Investigating alternative culturing protocols may

allow for a larger population of slugs, providing an equivalent samples size as insect qPCR studies.

The second RNAi construct, apoptosis inhibitor, also provided poor results, failing to work effectively based on lack of phenotype change. A dilemma with RNAi after no phenotype is seen is the choice between continuing an experiment by qPCR analysis or a functional assay, or trying another gene target. The apoptosis inhibitor gene was selected due to its success seen in insects within the research group, which included a clear phenotype. The lack of phenotype and unsuccessful qPCR experiments as well as the likelihood that apoptosis inhibitor would not be a legitimate target, due to conservation of the gene, resulted in the decision to not consider the use of the target further, but focus instead on another gene.

GAPDH was the final RNAi experiment discussed here. The lack of results in previous experiments focused the project on providing evidence that RNAi was working rather than considering long term potential of RNAi constructs. A HKG was chosen, with the express purpose of showing a gene knock-down effect with qPCR. HKG was specifically chosen due to reduce variability and highlight any knock down effect. This experiment also included considering longer time points, larger dsRNA quantities and pooling individuals to reduce variability. The inclusion of a second HKG was used in order to verify the first. It demonstrated that use of EFI-A showed a significant difference with GAPDH. As such the use of a secondary HKG for validation of the results was successful, but it furthered reduced confidence in the overall qPCR experimental results. The assumption that the work done with RNAi in insects would be transferable to molluscs was incorrect. An approach which prioritised identifying knock-down through a protein assay would have better validated RNAi.

Solid evidence of RNA interference working within *D. reticulatum* represented an important milestone, which was not met by this project. Future efforts may need to concentrate on initially proving the robustness of RNAi experiments in molluscs. Showing dsRNA traversal through the organism, uptake into the cell and presence over time are all experiments still to be done in *D. reticulatum* and would lend weight to future gene knock down experiments. Another area to expand is the

delivery method of RNAi, in this project experiments focused on injection assays, as a starting point to maximise the probability of dsRNA uptake. Feeding through incorporation of RNAi in feed or expression in plant species is a delivery method of interest and a more practical delivery method for downstream application.

Sequencing of the neural tissue generated a large selection of sequences not found in the digestive gland. However the lack of ion channels was potentially foreseeable with the same level of transcription not to be expected from ion channels as seen from the digestive enzymes. Despite this several sequences were of notable interest and their presence or absence in different tissues increased their notability. With RNAi having limited success, a protein based experiment was conducted. dTNF appeared to be present in both digestive gland and neural tissue and showed specific homology to ligands known to bind to TNF receptors, many of which initiate apoptosis. The results presented show only a limited assay to assess phenotype, which was unsuccessful. However the represents potential method for implementation of data produced in this work. It demonstrates that native proteins can be successfully identified, expressed, purified and assayed against the target organism. Future work could involve assessing yeast expression systems which are better able to express eukaryotic proteins which may require additional folding enzymes. However the potential that the protein caused a toxic effect when expressed, suggested by the lack of OD change in IPTG induced cultures should be taken into account. Whilst the focus of this project was to concentrate on generating an effect, producing toxins which are ubiquitously toxic to any organism will inhibit the viability of the protein as a downstream application.

A better target for protein expression may well be one of the most interesting sequences found in the dataset, the knottin-domain containing protein with close homology to spider venom toxins. This protein was found by chance when a list of all IPR terms was assessed by hand, toward the end of the project. The potential for this protein to be an ion channel blocker which spider toxins are imitating highlights it for further research. Presence of homologous proteins of unknown function in the insects suggests the results may also benefit research in other crop pests. Expression of this gene and its characterisation would be a recommended next step from this

project's work. Expression could follow the same protocol as that used for TNF. However spider toxins have been successfully expressed in yeast, if the protein is a toxin homologue then yeast may be a better expression system for assessing its potential.

Many of the aims of this project were successfully achieved. A mixture of molecular biology, bioinformatics, data analysis, RNAi and protein expression created a diverse set of work. However the overall target of taking the project from an organism with very little data available to a working vector causing lethality was overly-ambitious. Many experiments were left open-ended, with more data being required before useful conclusions could really be drawn. But the project contains a wealth of data available for this organism and identifies many starting points for future work. Initial experiments have been done which identify where many of the limitations of working with this organism lie and will hopefully better inform future research with *Deroceras reticulatum*.

Bibliography

Andersen, Lindbjerg, Jensen, and Ørntoft. 2004. 'Normalization of Real-Time Quantitative Reverse Transcription-PCR Data: A Model-Based Variance Estimation Approach to Identify Genes Suited for Normalization, Applied to Bladder and Colon Cancer Data Sets'. *Cancer Research* 64 (15): 5245 –5250. doi:10.1158/0008-5472.CAN-04-0496.

Andreasen. 1993. 'Metaldehyde Toxicosis in Ducklings'. *Journal of Veterinary Diagnostic Investigation* 5 (3) (July 1): 500–501. doi:10.1177/104063879300500341.

Badariotti, Fabien, Lelong, Dubos, and Favrel. 2007. 'Characterization of Chitinase-like Proteins (Cg-Clp1 and Cg-Clp2) Involved in Immune Defence of the Mollusc Crassostrea Gigas'. *FEBS Journal* 274 (14): 3646–3654. doi:10.1111/j.1742-4658.2007.05898.x.

Badariotti, Fabien, Thuau, Lelong, Dubos, and Favrel. 2007. 'Characterization of an Atypical Family 18 Chitinase from the Oyster Crassostrea Gigas: Evidence for a Role in Early Development and Immunity'. *Developmental & Comparative Immunology* 31 (6): 559–570. doi:10.1016/j.dci.2006.09.002.

Barbault, Landon, Guenneugues, Meyer, Schott, Dimarcq, and Vovelle. 2003. 'Solution Structure of Alo-3: a New Knottin-type Antifungal Peptide from the Insect Acrocinus Longimanus.' *Biochemistry* 42 (49) (December): 14434–14442.

Bardou, Leprince, Chichery, Vaudry, and Agin. 2010. 'Vasopressin/oxytocin-related Peptides Influence Long-term Memory of a Passive Avoidance Task in the Cuttlefish, Sepia Officinalis'. *Neurobiology of Learning and Memory* 93 (2) (February): 240–247. doi:10.1016/j.nlm.2009.10.004.

Barker. 2001. 'Gastropods on Land: Phylogeny, Diversity and Adaptive Morphology'. In *The Biology of Terrestrial Molluscs*, 565. Wallingford, Oxon: CABI Publishing.

Barker. 2002. *Molluscs as Crop Pests*. Wallingford, Oxon: CABI Publishing.

Barrett, Rawlings, and Woessner (Eds). 1998. Handbook of Proteolytic Enzymes. London: Academic Press

Bastien, Arnal, Bozonnet, Laguerre, Ferreira, Faure, Henrissat, et al. 2013. 'Mining for Hemicellulases in the Fungus-growing Termite Pseudacanthotermes Militaris Using Functional Metagenomics'. *Biotechnology for Biofuels* 6 (1): 78.

Berteau, McCort, Goasdoué, Tissot, and Daniel. 2002. 'Characterization of a New A-l-fucosidase Isolated from the Marine Mollusk Pecten Maximus That Catalyzes the Hydrolysis of A-l-fucose from Algal Fucoidan (Ascophyllum Nodosum)'. *Glycobiology* 12 (4) (April 1): 273–282. doi:10.1093/glycob/12.4.273.

Bleakley, Ferrie, Collum, and Burke. 2008. 'Self-poisoning with Metaldehyde'. *Emergency Medicine Journal* 25 (6) (June 1): 381–382. doi:10.1136/emj.2007.057414.

Bottke, Werner, Burschyk, and Volmer. 1988. 'On the Origin of the Yolk Protein Ferritin in Snails'. *Development Genes and Evolution* 197 (7): 377–382.

Boycott. 1934. 'The Habitats of Land Molusca in Britain'. *Journal of Ecology* (22): 1–38.

Breslow. 1979. 'Chemistry and Biology of the Neurophysins'. *Annual Review of Biochemistry* 48 (1) (June 1): 251–274. doi:10.1146/annurev.bi.48.070179.001343.

Brooks, Andrew, Crook, Wilcox, and Cook. 2003. 'A Laboratory Evaluation of the Palatability of Legumes to the Field Slug, Deroceras Reticulatum Müller'. *Pest Management Science* 59 (3): 245–251. doi:10.1002/ps.658.

Bruce, Fitches, Chougule, Bell, and Gatehouse. 2011. 'Recombinant Conotoxin, TxVIA, Produced in Yeast Has Insecticidal Activity'. *Toxicon* 58 (1) (July): 93–100. doi:doi: 10.1016/j.toxicon.2011.05.009.

Bibliography

Buckingham, Esmaeili, Wood, and Sattelle. 2004. 'RNA Interference: From Model Organisms Towards Therapy for Neural and Neuromuscular Disorders'. *Human Molecular Genetics* 13 (suppl 2) (October 1): R275–R288. doi:10.1093/hmg/ddh224.

Campbell. 2008. 'Metaldehyde Poisoning of Dogs'. *Veterinary Record* 163 (11) (September 13): 343–343. doi:10.1136/vr.163.11.343-a.

Campillos, Cases, Hentze, and Sanchez. 2010. 'SIREs: Searching for Iron-responsive Elements'. *Nucleic Acids Research* 38 (suppl 2) (July 1): W360 –W367. doi:10.1093/nar/gkq371.

Cantacessi, Jex, Hall, Young, Campbell, Joachim, Nolan, et al. 2010. 'A Practical, Bioinformatic Workflow System for Large Data Sets Generated by Next-generation Sequencing'. *Nucleic Acids Research* 38 (17): e171. doi:10.1093/nar/gkq667.

Carland and Gerwick. 2010. 'The C1q Domain Containing Proteins: Where Do They Come from and What Do They Do?' *Developmental & Comparative Immunology* 34 (8) (August): 785–790. doi:10.1016/j.dci.2010.02.014.

Celko. 2004. *Joe Celko's Trees and Hierarchies in SQL for Smarties, 2nd Edition*. 1st ed. Waltham, USA: Elsevier, Inc.

Cerutti, Mian, and Bateman. 2000. 'Domains in Gene Silencing and Cell Differentiation Proteins: The Novel PAZ Domain and Redefinition of the Piwi Domain'. *Trends in Biochemical Sciences* 25 (10) (October 1): 481–482. doi:10.1016/S0968-0004(00)01641-8.

Chapman. 2013. 'BioSQL Vs Own DB Schema for a Custom SNP + Annotations Data Storage - BioStar'. *BioSQL Vs Own DB Schema*. Accessed April 2. http://www.biostars.org/p/4892/.

Clark, Coward, Dawson, Henderson, and Martin. 1995. 'Metal Chelate Molluscicides: The Redistribution of Iron Diazaalkanolates from the Gut Lumen of the Slug, Deroceras Reticulatum (Müller) (Pulmonata: Limacidae)'. *Pesticide Science* 44 (4): 381–388. doi:10.1002/ps.2780440410.

Crichton and Charloteaux-Wauters. 1987. 'Iron Transport and Storage'. *European Journal of Biochemistry* 164 (3): 485–506. doi:10.1111/j.1432-1033.1987.tb11155.x.

Cruz, de Santos, Zafaralla, Ramilo, Zeikus, Gray, and Olivera. 1987. 'Invertebrate Vasopressin/oxytocin Homologs. Characterization of Peptides from Conus Geographus and Conus Straitus Venoms.' *Journal of Biological Chemistry* 262 (33) (November 25): 15821–15824.

Régis, Berteau, Jozefonvicz, and Goasdoue. 1999. 'Degradation of Algal (Ascophyllum Nodosum) Fucoidan by an Enzymatic Activity Contained in Digestive Glands of the Marine Mollusc Pecten Maximus'. *Carbohydrate Research* 322 (3–4) (December 12): 291–297. doi:10.1016/S0008-6215(99)00223-2.

Daquinag, Sato, Koda, Takao, Fukuda, Shimonishi, and Tsukamoto. 1999. 'A Novel Endogenous Inhibitor of Phenoloxidase from Musca Domestica Has a Cystine Motif Commonly Found in Snail and Spider Toxins†'. *Biochemistry* 38 (7) (January 29): 2179–2188. doi:10.1021/bi9819834.

Darling, Carey, and Feng. 2003. 'The Design, Implementation, and Evaluation of mpiBLAST'. In *4th International Conference on Linux Clusters: The HPC Revolution 2003 in Conjunction with ClusterWorld Conference & Expo, June 2003.*

Darl, Harrison, and Bottke. 1994. 'cDNA Cloning and Deduced Amino Acid Sequence of Two Ferritins: Soma Ferritin and Yolk Ferritin, from the Snail Lymnaea Stagnalis L.' *European Journal of Biochemistry* 222 (2): 353–366. doi:10.1111/j.1432-1033.1994.tb18874.x.

Davison and Blaxter. 2005. 'Ancient Origin of Glycosyl Hydrolase Family 9 Cellulase Genes'. *Molecular Biology and Evolution* 22 (5) (May): 1273 –1284. doi:10.1093/molbev/msi107.

Dekker, Voorn-Brouwer, Verhoek, Wennekes, Narayan, Speijer, Hollak, Overkleeft, Boot, and Aerts. 2011. 'The Cytosolic B-glucosidase GBA3 Does Not Influence Type 1 Gaucher Disease Manifestation'. *Gaucher Disease* 46 (1) (January 15): 19–26. doi:10.1016/j.bcmd.2010.07.009.

De Zoysa, Nikapitiya, Moon, Whang, Kim, and Lee. 2009. 'A Novel Fas Ligand in Mollusk Abalone: Molecular Characterization, Immune Responses and Biological Activity of the Recombinant Protein'. *Fish & Shellfish Immunology* 27 (3) (September): 423–432. doi:10.1016/j.fsi.2009.06.019.

Don, Cox, Wainwright, Baker, and Mattick. 1991. '"Touchdown" PCR to Circumvent Spurious Priming During Gene Amplification'. *Nucleic Acids Research* 19 (14) (July 25): 4008–4008. doi:10.1093/nar/19.14.4008.

Drickamer. 1992. 'Engineering Galactose-binding Activity into a C-type Mannose-binding Protein'. *Nature* 360 (6400) (November 12): 183–186. doi:10.1038/360183a0.

Edwards, Arancon, Vasko-Bennett, Little, and Askar. 2009. 'The Relative Toxicity of Metaldehyde and Iron Phosphate-based Molluscicides to Earthworms'. *Crop Protection* 28 (4) (April): 289–294. doi:10.1016/j.cropro.2008.11.009.

Eiberger, and Schilling. 2012. 'Cerebellins: Capstones to Bridge the Synaptic Cleft'. *Journal of Neurochemistry* 121 (5): 697–699. doi:10.1111/j.1471-4159.2012.07675.x.

Ester, van Rozen, and Molendijk. 2003. 'Field Experiments Using the Rhabditid Nematode Phasmarhabditis Hermaphrodita or Salt as Control Measures Against Slugs in Green Asparagus'. *Crop Protection* 22 (5) (June): 689–695. doi:10.1016/S0261-2194(03)00003-6.

Fabioux, Corporeau, Quillien, Favrel, and Huvet. 2009. 'In Vivo RNA Interference in Oyster –vasa Silencing Inhibits Germ Cell Development'. *FEBS Journal* 276 (9): 2566–2573. doi:10.1111/j.1742-4658.2009.06982.x.

Fainzilber, Gordon, Hasson, Spira, and Zlotkin. 1991. 'Mollusc-specific Toxins from the Venom of *Conus Textile Neovicarius*'. *European Journal of Biochemistry* 202 (2): 589–595.

Fairbairn, Cavallaro, Bernard, Mahalinga-Iyer, Graham, and Botella. 2007. 'Host-delivered RNAi: An Effective Strategy to Silence Genes in Plant Parasitic Nematodes'. *Planta* 226 (6) (November 1): 1525–1533. doi:10.1007/s00425-007-0588-x.

FAOSTAT United Nations. 2011. 'FAOSTAT Food and Agriculture Organization, United Nations'. http://faostat.fao.org/site/339/default.aspx.

Federhen. 2011. *NCBI Help Manuel - Entrez Taxonomy*. Bethesda (MD): National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/books/NBK53759/

Feng, Zhang, van Kesteren, Straub, van Nierop, Jin, Nejatbakhsh, et al. 2009. 'Transcriptome Analysis of the Central Nervous System of the Mollusc Lymnaea Stagnalis'. *BMC Genomics* 10 (1): 451.

Ferguson, Barratt, and Jones. 1988. 'Control of the Grey Field Slug (Deroceras Reticulatum (Muller)) by Stock Management Prior to Direct-drilled Pasture Establishment'. *The Journal of Agricultural Science* 111 (03): 443–449. doi:10.1017/S0021859600083611.

Fire, Xu, Montgomery, Kostas, Driver, and Mello. 1998. 'Potent and Specific Genetic Interference by Double-stranded RNA in Caenorhabditis Elegans'. *Nature* 391 (6669) (February 19): 806–811. doi:10.1038/35888.

Fretter and Peake. 1978. *Pulmonates - Volume 2A, Systematics, Evolution & Ecology*. 1st ed. London: Academic Press, Inc.

Fujii, Fang, Inoue, Murakami, and Sawayama. 2009. 'Enzymatic Hydrolyzing Performance of Acremonium Cellulolyticus and Trichoderma Reesei Against Three Lignocellulosic Materials'. *Biotechnology for Biofuels* 2 (1): 24.

Garthwaite, Barker, Parrish, and Smith. 2010. 'PESTICIDE USAGE SURVEY REPORT 235: ARABLE CROPS IN THE UNITED KINGDOM 2010'. 235. PESTICIDE USAGE SURVEY. Department for Food, Environment and Rural Affairs. http://www.fera.defra.gov.uk/scienceResearch/scienceCapabilities/landUseSustainability/surveys/documents/arable2010.pdf.

Gasteiger, Hoogland, Gattiker, Duvaud, Wilkins, Appel, and Bairoch. 2005. 'Protein Identification and Analysis Tools on the ExPASy Server'. In *The Proteomics Protocols Handbook*, edited by John M. Walker, 18th ed. Totowa, New Jersey: Humana Press.

Gavin, Banowertz, Griffith, Warrant, and Whittaker. 2007. 'Earthworms and Their Impact on Slug Control'. Experiment Station. Seed Production Research. Oregon State University: Department of Crop & Soil Science. http://www.ars.usda.gov/research/publications/publications.htm?seq_no_115=221850.

Gebler, Gilkes, Claeyssens, Wilson, Béguin, Wakarchuk, Kilburn, Miller, Warren, and Withers. 1992. 'Stereoselective Hydrolysis Catalyzed by Related Beta-1,4-glucanases and Beta-1,4-xylanases.' *Journal of Biological Chemistry* 267 (18) (June 25): 12559–12561.

Gerdol, Manfrin, De Moro, Figueras, Novoa, Venier, and Pallavicini. 2011. 'The C1q Domain Containing Proteins of the Mediterranean Mussel Mytilus Galloprovincialis: A Widespread and Diverse Family of Immune-related Molecules'. *Developmental & Comparative Immunology* 35 (6) (June): 635–643. doi:10.1016/j.dci.2011.01.018.

Glen. 2002. 'Integrated Control of Slug Damage'. *Pestic. Outlook* 13 (4): 137–141. doi:10.1039/B206510J.

Glen and Symondson. 2002. 'Influence of Soil Tillage on Slugs and Their Natural Enemies'. In *Soil Tillage in Agroecosystems*. Advances in Agroecology. CRC Press. http://dx.doi.org/10.1201/9781420040609.ch8.

Godan. 1999. *Molluscs - Their Significance for Science, Medicine, Commerce & Culture*. Berlin: Parey.

Gracy, Le-Nguyen, Gelly, Kaas, Heitz, and Chiche. 2008. 'KNOTTIN: The Knottin or Inhibitor Cystine Knot Scaffold in 2007'. *Nucleic Acids Research* 36 (suppl 1) (January 1): D314–D319. doi:10.1093/nar/gkm939.

Granato, Gonzales, Luz, Cassiola, Machado-Santelli, and Oliveira. 2005. 'Nop53p, an Essential Nucleolar Protein That Interacts with Nop17p and Nip7p, Is Required for pre-rRNA Processing in Saccharomyces Cerevisiae'. *FEBS Journal* 272 (17): 4450–4463.

Grisley and Boyle. 1990. 'Chitinase, a New Enzyme in Octopus Saliva'. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry* 95 (2): 311–316. doi:10.1016/0305-0491(90)90081-4.

Griss, Reisinger, Hermjakob, and Vizcaíno. 2012. 'jmzReader: A Java Parser Library to Process and Visualize Multiple Text and XML-based Mass Spectrometry Data Formats'. *PROTEOMICS* 12 (6): 795–798. doi:10.1002/pmic.201100578.

Haldane and Davis. 2009. 'Acute Toxicity in Five Dogs after Ingestion of a Commercial Snail and Slug Bait Containing Iron EDTA'. *Australian Veterinary Journal* 87 (July): 284–286. doi:10.1111/j.1751-0813.2009.00451.x.

Han and Zmasek. 2009. ‘phyloXML: XML for Evolutionary Biology and Comparative Genomics’. *BMC Bioinformatics* 10 (1): 356. doi:10.1186/1471-2105-10-356.

Harvey, Hrmova, De Gori, Varghese, and Fincher. 2000. ‘Comparative Modeling of the Three-dimensional Structures of Family 3 Glycoside Hydrolases’. *Proteins: Structure, Function, and Bioinformatics* 41 (2) (November 1): 257–269. doi:10.1002/1097-0134(20001101)41:2<257::AID-PROT100>3.0.CO;2-C.

He, Hou, Wang, and Zhao. 2012. ‘The Apoptosis Inhibitor Survivin Prevents Insect Midgut from Cell Death During Postembryonic Development’. *Molecular Biology Reports* 39 (2) (February 1): 1691–1699. doi:10.1007/s11033-011-0909-9.

Heinemann and Leipold. 2007. ‘Conotoxins of the O-superfamily Affecting Voltage-gated Sodium Channels’. *Cellular and Molecular Life Sciences* 64 (11) (June 6): 1329–1340. doi:10.1007/s00018-007-6565-5.

Henderson and Martin. 1990. ‘Control of Slugs with Contact-action Molluscicides’. *Annals of Applied Biology* 116 (2): 273–278. doi:10.1111/j.1744-7348.1990.tb06607.x.

Hodgkin and Huxley. 1952. ‘A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve’. *The Journal of Physiology* 117 (4) (August 28): 500–544.

Hollingsworth, Armstrong and Campbell. 2002. ‘Pest Control: Caffeine as a Repellent for Slugs and Snails’. *Nature* 417 (6892) (June 27): 915–916. doi:10.1038/417915a.

Holmgren. 1985. ‘Thioredoxin’. *Annual Review of Biochemistry* 54 (1) (June 1): 237–271. doi:10.1146/annurev.bi.54.070185.001321.

Howlett . 2005. ‘The Biology, Behaviour and Control of the Field Slug’. Doctor of Philosophy, The University of Newcastle Upon Tyne.

Huang Zhao Liu, Guan, Shi, Wang, Wu and He. 2012. ‘Gigabase-Scale Transcriptome Analysis on Four Species of Pearl Oysters’. *Marine Biotechnology* (September 1): 1–12. doi:10.1007/s10126-012-9484-x.

Igaki Kanda, Yamamoto-Goto, Kanuka, Kuranaga, Aigaki, and Miura. 2002. ‘Eiger, a TNF Superfamily Ligand That Triggers the Drosophila JNK Pathway’. *EMBO J* 21 (12) (June 17): 3009–3018. doi:10.1093/emboj/cdf306.

Ike, Isami, Tanabe, Nogawa, Ogasawara, Okada, and Morikawa. 2006. ‘Cloning and Heterologous Expression of the Exo-β-d-glucosaminidase-encoding Gene (gls93) from a Filamentous Fungus, Trichoderma Reesei PC-3-7’. *Applied Microbiology and Biotechnology* 72 (4) (October 1): 687–695. doi:10.1007/s00253-006-0320-y.

Isarankura and Runham. 1968. ‘Studies on the Replacement of the Gastropod Radula’. *Malacologia* 7: 71–91.

Iseli, Jongeneel and Bucher. 1999. ‘ESTScan: a Program for Detecting, Evaluating, and Reconstructing Potential Coding Regions in EST Sequences.’ *Proc Int Conf Intell Syst Mol Biol*: 138–48.

Jaubert-Possamai, Trionnaire, Bonhomme, Christophides, Rispe, and Tagu. 2007. ‘Gene Knockdown by RNAi in the Pea Aphid Acyrthosiphon Pisum’. *BMC Biotechnology* 7 (1): 63.

Jerala, Zerovnik, Kidric and Turk. 1998. ‘pH-induced Conformational Transitions of the Propeptide of Human Cathepsin L’. *Journal of Biological Chemistry* 273 (19) (May 8): 11498 –11504. doi:10.1074/jbc.273.19.11498.

Jiang, Loker and Zhang. 2006. ‘In Vivo and in Vitro Knockdown of FREP2 Gene Expression in the Snail Biomphalaria Glabrata Using RNA Interference’. *Developmental & Comparative Immunology* 30 (10): 855–866. doi:doi: DOI: 10.1016/j.dci.2005.12.004.

Katayama, Arakawa, Nakao, Ono, Aoki-Kinoshita, Yamamoto, Yamaguchi et al. 2010. ‘The DBCLS BioHackathon: Standardization and Interoperability for Bioinformatics Web Services and Workflows.’ *Journal of Biomedical Semantics* 1 (1): 8.

Katayama, Wilkinson, Micklem, Kawashima, Yamaguchi, Nakao, Yamamoto, et al. 2013. 'The 3rd DBCLS BioHackathon: Improving Life Science Data Integration with Semantic Web Technologies'. *Journal of Biomedical Semantics* 4 (1): 6.

Kazmierkiewicz, Czaplewski and Ciarkowski. 1997. 'Elucidation of Neurophysin/bioligand Interactions from Molecular Modeling'. *ACTA BIOCHIMICA POLONICA* 44 (3): 453–466.

Keymer, Gibson and Reynolds. 1991. 'Zoonoses and Other Findings in Hedgehogs (Erinaceus Europaeus): a Survey of Mortality and Review of the Literature'. *Veterinary Record* 128 (11) (March 16): 245–249.

Kibbe. 2007. 'OligoCalc: An Online Oligonucleotide Properties Calculator'. *Nucl. Acids Res.* 35 (suppl_2): W43–46.

King. 2007. 'Modulation of Insect Cav Channels by Peptidic Spider Toxins'. *Toxicon* 49 (4) (March 15): 513–530. doi:doi: DOI: 10.1016/j.toxicon.2006.11.012.

Korneev, Kemenes, Straub, Staras, Korneeva, Kemenes, Benjamin, and O'Shea. 2002. 'Suppression of Nitric Oxide (NO)-Dependent Behavior by Double-Stranded RNA-Mediated Silencing of a Neuronal NO Synthase Gene'. *The Journal of Neuroscience* 22 (11) (June 1): RC227.

Kumar and Blaxter. 2010. 'Comparing de Novo Assemblers for 454 Transcriptome Data'. *BMC Genomics* 11 (1): 571.

Laemmli. 1970. 'Cleavage of Structural Proteins During the Assembly of the Head of Bacteriophage T4'. *Nature* 227 (5259) (August 15): 680–685. doi:10.1038/227680a0.

Lau, Guiley, De, Potter, Carragher and MacRae. 2012. 'The Molecular Architecture of Human Dicer'. *Nat Struct Mol Biol* 19 (4) (April): 436–440. doi:10.1038/nsmb.2268.

Lee, Kim, Kim, Han, Lee, Lim, Chang, Kubo, and Kaang. 2001. 'Overexpression of and RNA Interference with the CCAAT Enhancer-Binding Protein on Long-Term Facilitation of Aplysia Sensory to Motor Synapses'. *Learning & Memory* 8 (4) (July 1): 220–226. doi:10.1101/lm.40201.

Leoni, Gallerani and Ceci. 2008. A genome walking strategy for the identification of eukaryotic nucleotide sequences adjacent to known regions. BIOTECHNIQUES 44, 229+.

Lemaire, Richardson, Goyer, Keryer, Lancelin, Makhatadze, and Jacquot. 2000. 'Primary Structure Determinants of the pH- and Temperature-dependent Aggregation of Thioredoxin'. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* 1476 (2) (February 9): 311–323. doi:10.1016/S0167-4838(99)00235-6.

Liang, Lange, Chen, van Oers, Vlak, and Westenberg. 2012. 'Functional Analysis of Two Inhibitor of Apoptosis (iap) Orthologs from Helicoverpa Armigera Nucleopolyhedrovirus'. *Virus Research* 165 (1) (April): 107–111. doi:10.1016/j.virusres.2012.01.012.

Li, Wang, Summer, Westfall, Brooks, Falk and Schrier. 2008. 'Molecular Mechanisms of Antidiuretic Effect of Oxytocin'. *Journal of the American Society of Nephrology* 19 (2) (February 1): 225–232. doi:10.1681/ASN.2007010029.

Liu, Li, Hong, Ni, Sheng and Shen. 2006. 'Cloning, Expression and Characterization of a Thermostable exo-β-D-glucosaminidase from the Hyperthermophilic Archaeon Pyrococcus Horikoshii'. *Biotechnology Letters* 28 (20) (October 1): 1655–1660. doi:10.1007/s10529-006-9137-0.

Long, Hamilton and Mitchell 2007. Asymmetric Competition via Induced Resistance: Specialist Herbivores Indirectly Suppress Generalist Preference and Populations. Ecology 88, 1232–1240.

Lu, Huang and Li. 1995. ‘Translocation of Ferritin and Biomineralization of Goethite in the Radula of the Limpet Cellana Toreuma Reeve’. *Experimental Cell Research* 219 (1) (July): 137–145. doi:10.1006/excr.1995.1214.

MacGregor, Janeček and Svensson. 2001. ‘Relationship of Sequence and Structure to Specificity in the A-amylase Family of Enzymes’. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology* 1546 (1) (March 9): 1–20. doi:10.1016/S0167-4838(00)00302-2.

Mackey. 2002. ‘Relational Modeling of Biological Data: Trees and Graphs - O’Reilly Media’. http://www.oreillynet.com/pub/a/network/2002/11/27/bioconf.html.

Marchler-Bauer, Lu, Anderson, Chitsaz, Derbyshire, DeWeese-Scott, Fong, et al. 2011. ‘CDD: a Conserved Domain Database for the Functional Annotation of Proteins’. *Nucleic Acids Research* 39 (suppl 1) (January 1): D225 –D229. doi:10.1093/nar/gkq1189.

Margulies, Egholm, Altman, Attiya, Bader, Bemben, Berka, et al. 2005. ‘Genome Sequencing in Microfabricated High-density Picolitre Reactors’. *Nature* 437 (7057): 376–380. doi:10.1038/nature03959.

Martinez, Patkaniowska, Urlaub, Lührmann, and Tuschl. 2002. ‘Single-Stranded Antisense siRNAs Guide Target RNA Cleavage in RNAi’. *Cell* 110 (5) (September 6): 563–574. doi:10.1016/S0092-8674(02)00908-X.

Matsuura, Kusunoki, Harada, and Kakudo. 1984. ‘Structure and Possible Catalytic Residues of Taka-Amylase A’. *Journal of Biochemistry* 95 (3) (January 1): 697–702.

May and Plasterk. 2005. ‘RNA Interference Spreading in C. Elegans’. In *Methods in Enzymology*, edited by and John J. Rossi David R. Engelke, Volume 392:308–315. Academic Press. http://www.sciencedirect.com/science/article/pii/S0076687904920186.

Mellanby. 1961. ‘Slugs at Low Temperatures’. *Nature* 189 (4768) (March 18): 944–944. doi:10.1038/189944b0.

“Metaldehyde Stewardship Group 22/10/12 - Breifing Notes: Autumn 2012.” 2012. http://www.getpelletwise.co.uk/uploads/literature/pdf/Briefing_Notes_-_Autumn2012-01.pdf.

Metcalf. 1971. ‘Structure - Activity Relationships for Insecticidal Carbamates’. *Bulletin of the World Health Organization* 44 (1-3): 43–78.

Middlebrooks, Pierce, and Bell. 2011. ‘Foraging Behavior Under Starvation Conditions Is Altered via Photosynthesis by the Marine Gastropod, Elysia Clarki’. *PLoS ONE* 6 (7) (July 20): e22162. doi:10.1371/journal.pone.0022162.

Milan, Coppe, Reinhardt, Cancela, Leite, Saavedra, Ciofi, et al. 2011. ‘Transcriptome Sequencing and Microarray Development for the Manila Clam, Ruditapes Philippinarum: Genomic Tools for Environmental Monitoring’. *BMC Genomics* 12 (1): 234.

Miller, Miyata, Brown and Tomoyasu. 2012. ‘Dissecting Systemic RNA Interference in the Red Flour Beetle Tribolium Castaneum: Parameters Affecting the Efficiency of RNAi’. *PLoS ONE* 7 (10) (October 25): e47431. doi:10.1371/journal.pone.0047431.

Mills. 2008. ‘Metaldehyde Poisoning of Dogs’. *Veterinary Record* 163 (10) (September 6): 310–310. doi:10.1136/vr.163.10.310-a.

Moccia, Cristo, Winlow and Cosmo. 2009. ‘GABAA- and AMPA-like Receptors Modulate the Activity of an Identified Neuron Within the Central Pattern Generator of the Pond Snail Lymnaea Stagnalis’. *Invertebrate Neuroscience* 9 (1) (March 1): 29–41. doi:10.1007/s10158-009-0086-x.

Moroz, Edwards, Puthanveettil, Kohn, Ha, Heyland, Knudsen, et al. 2006. ‘Neuronal Transcriptome of Aplysia: Neuronal Compartments and Circuitry’. *Cell* 127 (7) (December 29): 1453–1467. doi:doi: DOI: 10.1016/j.cell.2006.09.052.

Mundry, Bornberg-Bauer, Sammeth and Feulner. 2012. ‘Evaluating Characteristics of De Novo Assembly Software on 454 Transcriptome Data: A Simulation Approach’. *PLoS ONE* 7 (2) (February 27): e31410. doi:10.1371/journal.pone.0031410.

Mungall, Emmert and The FlyBase Consortium. 2007. ‘A Chado Case Study: An Ontology-based Modular Schema for Representing Genome-associated Biological Information’. *Bioinformatics* 23 (13) (July 1): i337–i346. doi:10.1093/bioinformatics/btm189.

Muñoz-Mérida, González-Plaza, Cañada, Blanco, García-López, Rodríguez, Pedrola, et al. 2013. ‘De Novo Assembly and Functional Annotation of the Olive (Olea Europaea) Transcriptome’. *DNA Research* (January 7). doi:10.1093/dnares/dss036. http://dnaresearch.oxfordjournals.org/content/early/2013/01/06/dnares.dss036.abstract.

Muragaki, Jacenko, Apte, Mattei, Ninomiya and Olsen. 1991. ‘The Alpha 2(VIII) Collagen Gene. A Novel Member of the Short Chain Collagen Family Located on the Human Chromosome 1.’ *The Journal of Biological Chemistry* 266 (12) (April): 7721–7727.

Nanjo and Sakai. 1990. ‘Purification and Characterization of an exo-beta-D-glucosaminidase, a Novel Type of Enzyme, from Nocardia Orientalis.’ *Journal of Biological Chemistry* 265 (17) (June 15): 10088–10094.

Narusuye, Kinugawa and Nagahama. 2005. ‘Responses of Cerebral GABA-containing CBM Neuron to Taste Stimulation with Seaweed Extracts in Aplysia Kurodai’. *Journal of Neurobiology* 65 (2): 146–156. doi:10.1002/neu.20182.

‘NCBI Help Manual - NCBI Bookshelf’. 2005. *NCBI Help Manual*. http://www.ncbi.nlm.nih.gov/books/NBK3831/.

Newton, Bullock, Hodgson. 2009. Glucosinolate polymorphism in wild cabbage (Brassica oleracea) influences the structure of herbivore communities. Oecologia 160, 63–76.

Nezlin and Voronezhskaya. 1997. ‘GABA-immunoreactive Neurones and Interactions of GABA with Serotonin and FMRFamide in a Peripheral Sensory Ganglion of the Pond Snail Lymnaea Stagnalis’. *Brain Research* 772 (1–2) (October 24): 217–225. doi:10.1016/S0006-8993(97)00835-4.

Nicholas. 1984. ‘The Biology of Reproduction in Two British Pulmonate Slugs’. Doctor of Philosophy, University of Wales.

Nielsen. 1962. ‘Carbohydrases in Soil and Litter Invertebrates’. *Oikos* 13 (2) (January 1): 200–215. doi:10.2307/3565085.

Nishimura, Furuno and Kato. 1988. ‘Biosynthesis and Processing of Lysosomal Cathepsin L in Primary Cultures of Rat Hepatocytes’. *Archives of Biochemistry and Biophysics* 263 (1) (May 15): 107–116. doi:10.1016/0003-9861(88)90618-2.

Numan and Bhosle. 2006. ‘α-l-Arabinofuranosidases: The Potential Applications in Biotechnology’. *Journal of Industrial Microbiology and Biotechnology* 33 (4) (April 1): 247–260. doi:10.1007/s10295-005-0072-1.

Leoni, Gallerani and Ceci. 2008. A genome walking strategy for the identification of eukaryotic nucleotide sequences adjacent to known regions. BIOTECHNIQUES 44, 229+.

Long, Hamilton and Mitchell. 2007. Asymmetric Competition via Induced Resistance: Specialist Herbivores Indirectly Suppress Generalist Preference and Populations. Ecology 88, 1232–1240.

Newton, Bullock and Hodgson 2009. Glucosinolate polymorphism in wild cabbage (Brassica oleracea) influences the structure of herbivore communities. Oecologia 160, 63–76.

Orians, Fritz, Hochwender, Albrectsen and Czesak 2013. How slug herbivory of juvenile hybrid willows alters chemistry, growth and subsequent susceptibility to diverse plant enemies. Annals of Botany 112, 757–765.

Wallingford, Liu and Zheng. 2010. Xenopus. Current Biology 20, R263–R264.

Osterauer, Marschner, Betz, Gerberding, Sawasdee, Cloetens, Haus, Sures, Triebskorn and Köhler. 2010. 'Turning Snails into Slugs: Induced Body Plan Changes and Formation of an Internal Shell'. *Evolution & Development* 12 (5): 474–483. doi:10.1111/j.1525-142X.2010.00433.x.

Pales Espinosa, Perrigault and Allam. 2010. 'Identification and Molecular Characterization of a Mucosal Lectin (MeML) from the Blue Mussel Mytilus Edulis and Its Potential Role in Particle Capture'. *Comparative Biochemistry and Physiology - Part A: Molecular & Integrative Physiology* 156 (4) (August): 495–501. doi:10.1016/j.cbpa.2010.04.004.

Pauchet, Wilkinson, van Munster, Augustin, Pauron and ffrench-Constant. 2009. 'Pyrosequencing of the Midgut Transcriptome of the Poplar Leaf Beetle Chrysomela Tremulae Reveals New Gene Families in Coleoptera'. *Insect Biochemistry and Molecular Biology* 39 (5-6) (May): 403–413. doi:10.1016/j.ibmb.2009.04.001.

Peng, Zha, He, Lu, Zhu, Han and He. 2011. 'Pyrosequencing the Midgut Transcriptome of the Brown Planthopper, Nilaparvata Lugens'. *Insect Molecular Biology* 20 (6): 745–762. doi:10.1111/j.1365-2583.2011.01104.x.

Pfaffl, Tichopad, Prgomet and Neuvians. 2004. 'Determination of Stable Housekeeping Genes, Differentially Regulated Target Genes and Sample Integrity: BestKeeper – Excel-based Tool Using Pair-wise Correlations'. *Biotechnology Letters* 26 (6) (March 1): 509–515. doi:10.1023/B:BILE.0000019559.84305.47.

Price and Gatehouse. 2008. 'RNAi-mediated Crop Protection Against Insects'. *Trends in Biotechnology* 26 (7) (July): 393–400. doi:doi: DOI: 10.1016/j.tibtech.2008.04.004.

Pyati, Bandani, Fitches, and Gatehouse. 2011. 'Protein Digestion in Cereal Aphids (Sitobion Avenae) as a Target for Plant Defence by Endogenous Proteinase Inhibitors'. *Journal of Insect Physiology* 57 (7) (July): 881–891. doi:10.1016/j.jinsphys.2011.03.024.

Pyati, Bell, Fitches, Price, Gatehouse and Gatehouse. 2009. 'Cathepsin L-like Cysteine Proteinase (DcCathL) from Delia Coarctata (wheat Bulb Fly): Basis of Insecticidal Activity'. *Insect Biochemistry and Molecular Biology* 39 (8) (August): 535–546. doi:10.1016/j.ibmb.2009.05.003.

Rae, Robertson and Wilson. 2009. 'Optimization of Biological (Phasmarhabditis Hermaphrodita) and Chemical (iron Phosphate and Metaldehyde) Slug Control'. *Crop Protection* 28 (9) (September): 765–773. doi:10.1016/j.cropro.2009.04.005.

Ren, Liu, Dong, Sun, Yang, Zhu and Jin. 2012. 'Evaluating de Bruijn Graph Assemblers on 454 Transcriptomic Data'. *PLoS ONE* 7 (12) (December 7): e51188. doi:10.1371/journal.pone.0051188.

Ronaghi, Karamohamed, Pettersson, Uhlen and Nyren. 1996. 'Real-Time DNA Sequencing Using Detection of Pyrophosphate Release'. *Analytical Biochemistry* 242 (1) (November 1): 84–89. doi:doi: DOI: 10.1006/abio.1996.0432.

Rothberg and Leamon. 2008. 'The Development and Impact of 454 Sequencing'. *Nat Biotech* 26 (10) (October): 1117–1124. doi:10.1038/nbt1485.

Runham. 1978. 'Reproduction and Its Control in Deroceras Reticulatum'. *Malacologia* 17 (341-350).

Runham and Hunter. 1970. *Terrestrial Slugs*. 1st ed. London: Hutchison & Co (publishers) LTD.

Ryan. 1978. 'Proteinase Inhibitors in Plant Leaves: A Biochemical Model for Pest-induced Natural Plant Protection'. *Trends in Biochemical Sciences* 3 (3) (July): 148–150. doi:10.1016/S0968-0004(78)90098-1.

Sadamoto, Takahashi,Okada, Kenmoku, Toyota, and Asakawa. 2012. 'De Novo Sequencing and Transcriptome Analysis of the Central Nervous System of Mollusc Lymnaea

Stagnalis by Deep RNA Sequencing'. *PLoS ONE* 7 (8) (August 1): e42546. doi:10.1371/journal.pone.0042546.

Sambrook and Russel. 2001. *Molecular Cloning: A Laboratory Manual*. 3rd ed. Cold Spring Habour Laboratory Press.

Sattelle and Buckingham. 2006. 'Invertebrate Studies and Their Ongoing Contributions to Neuroscience'. *Invertebrate Neuroscience* 6 (1) (March 1): 1–3. doi:10.1007/s10158-005-0014-7.

Scheltema and Schander. 2006. 'Exoskeletons: Tracing Molluscan Evolution'. *Venus (Tokyo)* 65 (1-2) (May): 19–26.

Schlüter, Benchabane, Munger, Kiggundu, Vorster, Goulet, Cloutier and Michaud. 2010. 'Recombinant Protease Inhibitors for Herbivore Pest Control: a Multitrophic Perspective'. *Journal of Experimental Botany* 61 (15) (October 1): 4169 –4183. doi:10.1093/jxb/erq166.

Schmidt, Schmid and Grossniklaus. 2012. 'Analysis of Plant Germline Development by High-throughput RNA Profiling: Technical Advances and New Insights'. *The Plant Journal* 70 (1) (April 1): 18–29. doi:10.1111/j.1365-313X.2012.04897.x.

Schöder, Port and Bennison. 2004. 'The Behavioural Response of Slugs and Snails to Novel Molluscicides, Irritants and Repellents'. *Pest Management Science* 60 (12): 1171–1177. doi:10.1002/ps.942.

Sen and Blau. 2006. "A Brief History of RNAi: The Silence of the Genes." *The FASEB Journal* 20 (9) (July 1): 1293–1299. doi:10.1096/fj.06-6014rev.

Sellar, Blake and Reid. 1991. 'Characterization and Organization of the Genes Encoding the A-, B- and C-chains of Human Complement Subcomponent C1q. The Complete Derived Amino Acid Sequence of Human C1q.' *The Biochemical Journal* 274 (Pt 2) (March): 481–490.

Shelton, Zhao and Roush. 2002. "ECONOMIC, ECOLOGICAL, FOOD SAFETY, AND SOCIAL CONSEQUENCES OF THE DEPLOYMENT OF BT TRANSGENIC PLANTS." *Annual Review of Entomology* 47 (1) (January 1): 845–881. doi:10.1146/annurev.ento.47.091201.145309

Shih, Chang, Chan, Chen, Chang, Tung, Deng and Yang. 2004. 'Acute Metaldehyde Poisoning in Taiwan.' *Veterinary and Human Toxicology* 46 (3) (June): 140–143.

Shindo and Van der Hoorn. 2008. 'Papain-like Cysteine Proteases: Key Players at Molecular Battlefields Employed by Both Plants and Their Invaders'. *Molecular Plant Pathology* 9 (1): 119–125. doi:10.1111/j.1364-3703.2007.00439.x.

Shi, Xie, Chen, Sun, Zhou, Liu, Gao, Kyrpides, No and Yuan. 2013. 'Comparative Genomic Analysis of the Endosymbionts of Herbivorous Insects Reveals Eco-Environmental Adaptations: Biotechnology Applications'. *PLoS Genet* 9 (1) (January 10): e1003131. doi:10.1371/journal.pgen.1003131.

Siegfried, Waterfield and Ffrench-Constant. 2005. 'Expressed Sequence Tags from Diabrotica Virgifera Virgifera Midgut Identify a Coleopteran Cadherin and a Diversity of Cathepsins'. *Insect Molecular Biology* 14 (2): 137–143. doi:10.1111/j.1365-2583.2005.00538.x.

Sifuentes-Romero, Milton and García-Gasca. 2011. 'Post-transcriptional Gene Silencing by RNA Interference in Non-mammalian Vertebrate Systems: Where Do We Stand?' *Mutation Research/Reviews in Mutation Research* 728 (3) (November): 158–171. doi:10.1016/j.mrrev.2011.09.001.

Sijen, Fleenor, Simmer, Thijssen, Parrish, Timmons, Plasterk and Fire. 2001. 'On the Role of RNA Amplification in dsRNA-Triggered Gene Silencing'. *Cell* 107 (4) (November 16): 465–476. doi:10.1016/S0092-8674(01)00576-1.

Silver, Best, Jiang, and Thein. 2006. 'Selection of Housekeeping Genes for Gene Expression Studies in Human Reticulocytes Using Real-time PCR'. *BMC Molecular Biology* 7 (1): 33.

Simms, Dawson, Paton and Wilson. 2006. 'Identification of Environmental Factors Limiting Plant Uptake of Metaldehyde Seed Treatments Under Field Conditions'. *Journal of Agricultural and Food Chemistry* 54 (10) (April 13): 3646–3650. doi:10.1021/jf060231a.

Simms,Mullins and Wilson. 2002. 'Seed Dressings to Control Slug Damage in Oilseed Rape'. *Pest Management Science* 58 (7): 687–694. doi:10.1002/ps.514.

Sirakov, Zarrella, Borra, Rizzo, Biffali, Arnone and Fiorito. 2009. 'Selection and Validation of a Set of Reliable Reference Genes for Quantitative RT-PCR Studies in the Brain of the Cephalopod Mollusc Octopus Vulgaris'. *BMC Molecular Biology* 10 (1): 70.

Smith, Tachibana, Pohl, Lee, Thanarajasingam, Portier, Ueki, et al. 2000. 'A Transcript Map of the Chromosome 19q-Arm Glioma Tumor Suppressor Region'. *Genomics* 64 (1) (February 15): 44–50. doi:10.1006/geno.1999.6101.

Solem. 1974. *The Shell Makers: Introducing Mollusks*. New York, USA: John Wiley & Sons Inc.

Solis, Santi-Rocca, Perdomo, Weber and Guillén. 2009. 'Use of Bacterially Expressed dsRNA to Downregulate Entamoeba Histolytica Gene Expression'. *PLoS ONE* 4 (12) (December 23): e8424. doi:10.1371/journal.pone.0008424.

Sollas. 1907. "The Molluscan Radula: Its Chemical Composition, and Some Points in Its Development." *Quarterly Journal of Microscopical Science* S2-51 (201) (February): 115–136.

Sparks, Quistad, Cole and Casida. 1996. 'Metaldehyde Molluscicide Action in Mice: Distribution, Metabolism, and Possible Relation to GABAergic System'. *Pesticide Biochemistry and Physiology* 55 (3) (July): 226–236. doi:10.1006/pest.1996.0052.

Speiser and Kistler. 2002. 'Field Tests with a Molluscicide Containing Iron Phosphate'. *Crop Protection* 21 (5) (June): 389–394. doi:10.1016/S0261-2194(01)00120-X.

Stam, Danchin, Rancurel, Coutinho and Henrissat. 2006. 'Dividing the Large Glycoside Hydrolase Family 13 into Subfamilies: Towards Improved Functional Annotations of A-amylase-related Proteins'. *Protein Engineering Design and Selection* 19 (12) (December 1): 555–562. doi:10.1093/protein/gzl044.

Stone, B. A., and J. E. Morton. 1958. 'The Distribution of Cellulases and Related Enzymes in Mollusca'. *Journal of Molluscan Studies* 33 (3) (December 1): 127–141.

Studier, F. William. 2005. 'Protein Production by Auto-induction in High-density Shaking Cultures'. *Protein Expression and Purification* 41 (1) (May): 207–234. doi:10.1016/j.pep.2005.01.016.

Sultan, Marc, Marcel H. Schulz, Hugues Richard, Alon Magen, Andreas Klingenhoff, Matthias Scherf, Martin Seifert, et al. 2008. 'A Global View of Gene Activity and Alternative Splicing by Deep Sequencing of the Human Transcriptome'. *Science* 321 (5891) (August 15): 956–960. doi:10.1126/science.1160342.

Sydorskyy, Dilworth, Halloran, Yi, Makhnevych, Wozniak and Aitchison. 2005. 'Nop53p Is a Novel Nucleolar 60S Ribosomal Subunit Biogenesis Protein.' *Biochem. J.* 388 (3): 819–826.

Tao and Fletcher. 2013. 'Metaldehyde Removal from Aqueous Solution by Adsorption and Ion Exchange Mechanisms onto Activated Carbon and Polymeric Sorbents'. *Journal of Hazardous Materials* 244–245 (0) (January 15): 240–250. doi:10.1016/j.jhazmat.2012.11.014.

Tao, Stearns, Dong, Wu and Sahagian. 1994. 'The Proregion of Cathepsin L Is Required for Proper Folding, Stability, and ER Exit'. *Archives of Biochemistry and Biophysics* 311 (1) (May): 19–27. doi:10.1006/abbi.1994.1203.

Takeuchi, Kawashima, Koyanagi, Gyoja, Tanaka, Ikuta, Shoguchi, et al. 2012. "Draft Genome of the Pearl Oyster Pinctada Fucata: A Platform for Understanding Bivalve Biology." *DNA Research* 19 (2) (April 1): 117–130. doi:10.1093/dnares/dss005

Tomoyasu, Miller, Tomita, Schoppmeier, Grossmann and Bucher. 2008. 'Exploring Systemic RNA Interference in Insects: a Genome-wide Survey for RNAi Genes in Tribolium'. *Genome Biology* 9 (1): R10.

Triebskorn. 1989. 'Ultrastructural Changes in Digestive System of Deroceras Reticulatum (Müller) Induced by a Carbamate Molluscicide and by Metaldehyde'. *Malacologia* 31 (1): 141–156.

Tulli, Carmona, López, Manetti, Vincini and Cendoya. 2009. 'Predation on the Slug Deroceras Reticulatum (Pulmonata: Stylommatophora) by Scarites Anthracinus (Coleoptera: Carabidae)'. *Ecologia Austral* 19 (April): 55–61.

Urade, Oberdick, Molinar-Rode and Morgan. 1991. 'Precerebellin Is a Cerebellum-specific Protein with Similarity to the Globular Domain of Complement C1q B Chain'. *Proceedings of the National Academy of Sciences* 88 (3) (February 1): 1069–1073.

Valentine, Rumbeiha, Hensley and Halse. 2007. 'Arsenic and Metaldehyde Toxicosis in a Beef Herd'. *Journal of Veterinary Diagnostic Investigation* 19 (2) (March 1): 212–215. doi:10.1177/104063870701900216.

Van Dam, Horn, Mareš and Baldwin. 2001. 'Ontogeny Constrains Systemic Protease Inhibitor Response in Nicotiana Attenuata'. *Journal of Chemical Ecology* 27 (3) (March 1): 547–568. doi:10.1023/A:1010341022761.

Vandesompele, De Preter, Pattyn, Poppe, Van Roy, De Paepe and Speleman. 2002. 'Accurate Normalization of Real-time Quantitative RT-PCR Data by Geometric Averaging of Multiple Internal Control Genes'. *Genome Biology* 3 (7): research0034.1–research0034.11. doi:10.1186/gb-2002-3-7-research0034.

Van Kesteren, Smit, De Lange, Kits, Van Golen, Van Der Schors, De With, Burke and Geraerts. 1995. 'Structural and Functional Evolution of the Vasopressin/oxytocin Superfamily: Vasopressin-related Conopressin Is the Only Member Present in Lymnaea, and Is Involved in the Control of Sexual Behavior'. *The Journal of Neuroscience* 15 (9) (September 1): 5989–5998.

Van Kesteren, Smit, Dirks, de With, Geraerts and Joosse. 1992. 'Evolution of the Vasopressin/oxytocin Superfamily: Characterization of a cDNA Encoding a Vasopressin-related Precursor, Preproconopressin, from the Mollusc Lymnaea Stagnalis.' *Proceedings of the National Academy of Sciences* 89 (10) (May 15): 4593–4597. doi:10.1073/pnas.89.10.4593.

Vouzis and Sahinidis. 2011. 'GPU-BLAST: Using Graphics Processors to Accelerate Protein Sequence Alignment'. *Bioinformatics* 27 (2) (January 15): 182–188. doi:10.1093/bioinformatics/btq644.

Walker, Glen and Shewry. 1998. 'Purification and Characterization of a Digestive Cysteine Proteinase from the Field Slug (Deroceras Reticulatum): A Potential Target for Slug Control'. *Journal of Agricultural and Food Chemistry* 46 (7) (July 1): 2873–2881. doi:10.1021/jf980082z.

Walker. 1972. 'The Digestive System of the Slug, Agriolimax Reticulatus (Müller): Experiments on Phagocytosis and Nutrient Absorption'. *Journal of Molluscan Studies* 40 (1) (April 1): 33 –43.

Wallingford, Liu, Zheng. 2010. Xenopus. Current Biology 20, R263–R264.

Wang, Wang, Zhang, Zhou, Siva and Song. 2012. 'A C1q Domain Containing Protein from Scallop Chlamys Farreri Serving as Pattern Recognition Receptor with Heat-Aggregated IgG Binding Activity'. *PLoS ONE* 7 (8) (August 15): e43289. doi:10.1371/journal.pone.0043289.

Wang, Yang, Zhou, Qiu, Wang, Zhang, Gao, et al. 2011. 'A Primitive Toll-like Receptor Signaling Pathway in Mollusk Zhikong Scallop Chlamys Farreri'. *Developmental & Comparative Immunology* 35 (4) (April): 511–520. doi:doi: DOI: 10.1016/j.dci.2010.12.005.

Wiederanders and Kirschke. 1989. 'The Processing of a Cathepsin L Precursor in Vitro'. *Archives of Biochemistry and Biophysics* 272 (2) (August 1): 516–521. doi:10.1016/0003-9861(89)90247-6.

Wilkinson. 2012. 'Exact and Approximate Area-Proportional Circular Venn and Euler Diagrams'. *Visualization and Computer Graphics, IEEE Transactions On* 18 (2) (February): 321–331. doi:10.1109/TVCG.2011.56.

Windley, Herzig, Dziemborowicz, Hardy, King, and Nicholson. 2012. 'Spider-Venom Peptides as Bioinsecticides'. *Toxins* 4 (3): 191–227. doi:10.3390/toxins4030191.

Wood, Miljenovic, Cai, Raven, Kaas, Escoubas, Herzig, Wilson and King. 2009. 'ArachnoServer: a Database of Protein Toxins from Spiders'. *BMC Genomics* 10 (1): 375.

'Woodstox'. 2013. *Woodstox High-performance XML Processor*. Accessed June 5. http://woodstox.codehaus.org/.

Xie, Xiao, Chen, Xu and Zhang. 2012. 'miRDeepFinder: a miRNA Analysis Tool for Deep Sequencing of Plant Small RNAs'. *Plant Molecular Biology* 80 (1) (September 1): 75–84. doi:10.1007/s11103-012-9885-2.

Yamaura, Takahashi and Suzuki. 2008. 'Identification and Tissue Expression Analysis of C-type Lectin and Galectin in the Pacific Oyster, Crassostrea Gigas'. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* 149 (1) (January): 168–175. doi:10.1016/j.cbpb.2007.09.004.

Yim, Kim, Ko, Cho, Kim, Kim, Lee and Park. 2007. 'The Putative Tumor Suppressor Gene GLTSCR2 Induces PTEN-modulated Cell Death'. *Cell Death Differ* 14 (11) (July 27): 1872–1879.

Young and Armstrong. 2001. 'Slugs, Snails and Iron Based Baits: An Increasing Problem and a Low Toxic Specific Action Solution'. In *Proceedings of 10th Agronomy Conference 2001*. Hobart, Tasmania.

Yuasa, Furuta, Nakamura and Takagi. 1998. 'Cloning and Sequencing of Three C-type Lectins from Body Surface Mucus of the Land Slug, Incilaria Fruhstorferi'. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* 119 (3) (March): 479–484. doi:10.1016/S0305-0491(98)00008-X.

Zdobnov and Apweiler. 2001. 'InterProScan – an Integration Platform for the Signature-recognition Methods in InterPro'. *Bioinformatics* 17 (9): 847 –848. doi:10.1093/bioinformatics/17.9.847.

Zelensky and Gready. 2005. 'The C-type Lectin-like Domain Superfamily'. *FEBS Journal* 272 (24): 6179–6217. doi:10.1111/j.1742-4658.2005.05031.x.

Zhang, Fang, Guo, Li, Luo, Xu, Yang, et al. 2012. 'The Oyster Genome Reveals Stress Adaptation and Complexity of Shell Formation'. *Nature* 490 (7418) (October 4): 49–54. doi:10.1038/nature11413.

Zhang, Meng, Jiang, Wang, Xie and Zhang. 2003. 'A Novel Ferritin Subunit Involved in Shell Formation from the Pearl Oyster (Pinctada Fucata)'. *Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* 135 (1) (May): 43–54. doi:10.1016/S1096-4959(03)00050-2.

Zhang, Zhou, Liu, Cao, He, Huo, Qin, Yao and Ringø. 2013. 'High-yield Production of a Chitinase from Aeromonas Veronii B565 as a Potential Feed Supplement for Warm-water Aquaculture'. *Applied Microbiology and Biotechnology* (June 18): 1–12. doi:10.1007/s00253-013-5023-6.