P-GLYCOPROTEIN GENES IN
HAEMONCHUS CONTORTUS

JAMES D. W. KENWORTHY

A thesis submitted for the degree of Master of Philosophy

University of Bath

Department of Biology and Biochemistry

May 2013

COPYRIGHT

Attention is drawn to the fact that copyright of this thesis rests with its author. A copy of this thesis has been supplied on condition that anyone who consults it is understood to recognise that its copyright rests with its author and they must not copy it or use material from it except as permitted by law or with the consent of the author.

This thesis may be made available for consultation within the University Library and may be photocopied or lent to other libraries for the purpose of consultation.
ACKNOWLEDGEMENTS

Thanks to my supervisors

Dr Priest, Dr Wolstenholme, Dr Walsh and Dr Coles.

Thanks for all the support from my family and friends.
TABLE OF CONTENTS

Acknowledgements ........................................................................................................2
Table of Contents ............................................................................................................3
Table of Figures ...............................................................................................................4
Table of Tables ...............................................................................................................5
Table of Sequences ........................................................................................................5
Table of Equations .........................................................................................................5
List of Abbreviations ......................................................................................................6
Summary .........................................................................................................................7
1 Introduction ................................................................................................................8
  1.1 Helminths ................................................................................................................8
  1.2 Anthelmintics .........................................................................................................11
  1.3 Anthelmintic Resistance .......................................................................................12
  1.4 Genetics of Anthelmintic Resistance ................................................................13
  1.5 Study of Genetics of Anthelmintic Resistance .....................................................18
  1.6 Aim of Project .......................................................................................................22
2 Materials and methods ................................................................................................23
  2.1 Bioinformatics .......................................................................................................23
  2.1.1 Alignments along Entire Gene ........................................................................23
  2.1.1.1 Alignments from Assembled Contigs and Expressed Sequence Tags
       Databases .............................................................................................................27
  2.1.1.2 Alignments from Eight Databases ................................................................28
  2.1.2 Alignments to Specific Domains ......................................................................29
  2.1.2.1 Alignments to Inter-nucleotide Binding Domains ......................................29
  2.1.2.2 Alignments to ATP Binding Cassettes .......................................................30
  2.1.2.3 Alignments to Both ATP Binding Cassettes and Transmembrane
       Domains ..............................................................................................................31
  2.1.3 Alignments along Entire Gene ........................................................................32
  2.1.3.1 Gene Structure Prediction Algorithms .....................................................32
3 Results .........................................................................................................................37
  3.1 Alignments along Entire Gene .............................................................................37
  3.1.1 Alignments from Assembled Contigs and Expressed Sequence Tags
       Databases .............................................................................................................37
  3.1.2 Alignments from Eight Databases ....................................................................39
  3.2 Alignments to Specific Domains .........................................................................41
  3.2.1 Alignments to ATP Binding Cassettes ............................................................42
  3.2.1.1 Alignments to Both ATP Binding Cassettes and Transmembrane
       Domains ..............................................................................................................47
3.3 Alignments along Entire Gene .............................................................................50
  3.3.1 Gene Structure Prediction Algorithms ............................................................57
4 Discussion ....................................................................................................................69
Bibliography ..................................................................................................................79
Appendix ........................................................................................................................85
TABLE OF FIGURES

Figure 1: Alignments of *H. contortus* Sequences from CON Database with *Cel*-PGP-4. .................................................................38

Figure 2: Alignments of *H. contortus* Sequences from Multiple Databases with Homologous *Cel*-PGP. .................................................................40

Figure 3: Motif and Domain Alignments from CD-Search to a *H. contortus* contig from ASS Database.................................................................42

Figure 4: Alignment of a *H. contortus* contig from ASS database with PSSM of ABC model. .........................................................................................42

Figure 5: Alignment of a *H. contortus* contig from ASS database with ABC Transporter Motif in ABC Domain Model.................................................43

Figure 6: Alignments of *H. contortus* contigs from ASS database with ABC Domain Model. .........................................................................................43

Figure 7: Neighbour-Joining Tree of a *H. contortus* contig from ASS Database with the Aligned Portion of ABC Domain from *Cel*-PGP........45

Figure 8: Alignment of *H. contortus* contigs from ASS Database Aligned with 5' portion of ABC Domain Model. .................................................................46

Figure 9: Coverage of *Hco*-PGP–2.1.A to ABC and TMD Domain Models. .....47

Figure 10: Alignment of *H. contortus* contigs from ASS Database at the Intersection of TMD and ABC Domains.................................................................48

Figure 11: Alignments of *H. contortus* contigs from C09 Database with Homologous *Cel*-PGP. .........................................................................................49

Figure 12: Dot Plot of *H. contortus* supercontig from S09 Database Aligned with *Cel*-PGP-4. .........................................................................................50

Figure 13: Alignments of *H. contortus* supercontigs from S09 Database with Homologous *Cel*-PGP. .........................................................................................65
TABLE OF TABLES
Table 1: Taxonomy of *Haemonchus contortus* .................................................................9
Table 2: *Haemonchus* Species ..........................................................................................11
Table 3: Evidence of the Genetics of IVM Resistance ............................................................17
Table 4: Approaches to Studying Genetics and Genomics ....................................................18
Table 5: Evidence for the Number of *pgp* Genes in Different Databases .........................41
Table 6: Evidence for the Number of *pgp* Genes in ASS Database .................................48
Table 7: Evidence for the Number of *pgp* Genes in C09 Database .................................49
Table 8: Summarised Alignments of *H. contortus supercontigs* from S09
                  Database with Homologous *C. elegans* – PGP .................................................67
Table 9: Number of *pgp* Genes in *C. elegans* and *H. contortus* .............................68

TABLE OF SEQUENCES
Sequence 1: Annotated *H. contortus supercontig* 0001880 from S09 Database.57
Sequence 2: Annotated *H. contortus supercontig* 0001880 from S09 Database
                  using the GeneWise Algorithm ...........................................................................63

TABLE OF EQUATIONS
Equation 1: E-value for Alignments. ....................................................................................27
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>ATP binding cassette</td>
</tr>
<tr>
<td>ASS</td>
<td>assembled contigs database released in November 2007</td>
</tr>
<tr>
<td>ATP</td>
<td>adenosine triphosphate</td>
</tr>
<tr>
<td>AV</td>
<td>avermectins</td>
</tr>
<tr>
<td>BAC</td>
<td>bacterial artificial chromosome</td>
</tr>
<tr>
<td>BAC</td>
<td>BAC end sequence reads database</td>
</tr>
<tr>
<td>BLAST</td>
<td>Basic Local Alignment Search Tool</td>
</tr>
<tr>
<td>BLASTP</td>
<td>BLAST for amino acid queries of amino acid subjects</td>
</tr>
<tr>
<td>BLASTX</td>
<td>BLAST for nucleotide queries of amino acid subjects</td>
</tr>
<tr>
<td>C09</td>
<td>assembled contigs database released in August 2009</td>
</tr>
<tr>
<td>CD</td>
<td>Conserved Domain</td>
</tr>
<tr>
<td>cd</td>
<td>conserved domain</td>
</tr>
<tr>
<td>CDD</td>
<td>Conserved Domains Database</td>
</tr>
<tr>
<td>COG</td>
<td>clusters of orthologous groups</td>
</tr>
<tr>
<td>CON</td>
<td>assembled contigs database released in January 2006</td>
</tr>
<tr>
<td>contigs</td>
<td>contiguous sequences derived from shotgun sequencing</td>
</tr>
<tr>
<td>C-terminus</td>
<td>carboxyl-terminus</td>
</tr>
<tr>
<td>EST</td>
<td>expressed sequence tags</td>
</tr>
<tr>
<td>EST</td>
<td>expressed sequence tags database</td>
</tr>
<tr>
<td>E-value</td>
<td>Expected value</td>
</tr>
<tr>
<td>FECRT</td>
<td>Faecal Egg Count Reduction Test</td>
</tr>
<tr>
<td>FOS</td>
<td>fosmid end sequence reads database</td>
</tr>
<tr>
<td>gDNA</td>
<td>genomic DNA</td>
</tr>
<tr>
<td>HSPS</td>
<td>high-scoring segment pairs</td>
</tr>
<tr>
<td>IBD</td>
<td>inter-nucleotide binding domains</td>
</tr>
<tr>
<td>IVM</td>
<td>ivermectin</td>
</tr>
<tr>
<td>kb</td>
<td>kilobase pairs</td>
</tr>
<tr>
<td>MI</td>
<td>milbemycins</td>
</tr>
<tr>
<td>ML</td>
<td>macrocyclic lactones</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger ribonucleic acid</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>nr</td>
<td>non-redundant protein sequences</td>
</tr>
<tr>
<td>N-terminus</td>
<td>amino-terminus</td>
</tr>
<tr>
<td>pgp</td>
<td>P-gp gene</td>
</tr>
<tr>
<td>PGP</td>
<td>P-gp protein</td>
</tr>
<tr>
<td>P-gp</td>
<td>P-glycoprotein</td>
</tr>
<tr>
<td>PHYLIP</td>
<td>Phylogeny Inference Package</td>
</tr>
<tr>
<td>PSSM</td>
<td>Position-Specific Scoring Matrix</td>
</tr>
<tr>
<td>qPCR</td>
<td>quantitative real-time polymerase chain reaction</td>
</tr>
<tr>
<td>RPS-BLAST</td>
<td>Reverse Position-Specific Iterative BLAST for queries of PSSM</td>
</tr>
<tr>
<td>S</td>
<td>alignment score</td>
</tr>
<tr>
<td>S09</td>
<td>supercontigs database released in August 2009</td>
</tr>
<tr>
<td>Sanger</td>
<td>Wellcome Trust Sanger Institute</td>
</tr>
<tr>
<td>SEQ1</td>
<td>sequence reads database released in December 2004</td>
</tr>
<tr>
<td>SEQ2</td>
<td>sequence reads database released in August 2005</td>
</tr>
<tr>
<td>SEQ3</td>
<td>sequence reads database released in November 2005</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphisms</td>
</tr>
<tr>
<td>TBLASTN</td>
<td>BLAST for amino acid queries of nucleotide subjects</td>
</tr>
<tr>
<td>TMD</td>
<td>transmembrane domain</td>
</tr>
<tr>
<td>UNA</td>
<td>unassembled reads database released in January 2006</td>
</tr>
<tr>
<td>WU-BLAST</td>
<td>Washington University Basic Local Alignment Search Tool</td>
</tr>
</tbody>
</table>
SUMMARY

Control of *Haemonchus contortus* currently relies on anthelmintics including ivermectin (IVM). The existence of IVM resistance threatens the sheep industry. Detailed knowledge of the underlying genetic mechanisms may reveal opportunities to avoid selection of IVM resistance. Molecular tests are required in order to more sensitively identify and quantify resistance and then test competing hypotheses on the prevention and control of resistance. It is first necessary to identify those genes or loci under selection following IVM application and detect genotypic markers of IVM resistance.

However, as the genetic basis of IVM resistance is not yet established and in the absence of genomic tools currently available for *H. contortus*, the approach taken to date is the analysis of the gene families implicated in IVM resistance including the P-glycoproteins (P-gps). *H. contortus* has an incomplete genome without annotation and only one P-gp messenger RNA (mRNA) transcript has been fully sequenced. The aim of this project was to discover the incomplete and dispersed P-gp genetic sequences in the sequence databases and construct the putative genes.

Bioinformatics was used to reveal *H. contortus* sequences that map both from and to known P-gp sequences in *Caenorhabditis elegans* by a process of Basic Local Alignment Search Tool (BLAST) algorithms, protein distance matrices, phylogenetics and gene structure prediction algorithms. Evidence of up to 11 P-gp genes (*pgp*) was discovered, which is fewer than the 15 in *C. elegans*. There are eight clades in *C. elegans* (*Cel-pgp*-1; 2; 3/4; 5/6/7/8; 9; 10; 11; 12/13/14/15) of which seven have representatives in *H. contortus*; there are no homologues to the entire clade of four genes *Cel-pgp*-5 to 8 inclusive. All the duplications in *C. elegans* are on the X chromosome but none of these groups has more than one homologue in *H. contortus*. Conversely autosomal *Cel-pgp*-2 and *Cel-pgp*-9 with evidence of multiple homologues in *H. contortus* are not duplicated in *C. elegans*.
1  INTRODUCTION

1.1  Helminths

Helminths comprise trematodes (flat worms), cestodes (tape worms) and nematodes (roundworms). The Nematoda phylum (Diesing, 1861) as roundworms use their longitudinal muscles and high internal pressure to provide the thrashing motion of their entire body for locomotion. Larval and adult motility assays utilise this action to measure anthelmintic inhibition in sensitive organisms and test for successful movement even in the presence of anthelmintic in resistant nematodes.

Reproduction is sexual in the majority of nematodes and in the two organisms in this study. The varied progeny increases the chance in each generation for some individuals to be resistant or less sensitive to noxious stimuli such as anthelmintic. Further generations of meiosis, recombination and crosses between male and females that are resistant or less sensitive enables a Nematoda population to become resistant.

The Nematoda are abundant in terms of species but importantly for this study also have large population sizes in an environment such as thousands in one ruminant host. This opportunity for multiple sexual partners further enhances variability across the population in a host or environment. Therefore if a novel or intermittent challenge such as anthelmintic application is less that 100% efficacious the surviving population will be large and concentrated enough to reproduce successfully. The nematodes are noted for their wide range of habitats, some species are very specific and limited whilst others are capable of a variety of environments. The parasitic nematode in this work has a global range and is known to infect a multitude of hosts.

Nematodes are protected by a secreted cuticle that is sufficient protection for invading the digestive tracts in the case of parasites. The pathogenic worm in question is safely able to pass through the preceding three stomachs to burrow into and then return into the lumen of the acidic fourth stomach of ruminant hosts. This cuticle layer defines the life cycle as four moults occur in the egg and larval stages before the adult stage in both the model organism and pathogen.(1)
One of the important sense organs of Nematoda are papillae, these are chemoreceptors\(^{(1)}\) and provide the input for moving away from a noxious chemical such as anthelmintic. This ability is used in an adult motility assay through a barrier away from the anthelmintic under investigation. The amphid organs and their internal molecular biology are heavily implicated in resistance to two main classes of anthelmintics. A particular feature of the Nematoda nervous system is a ring of neurones encircling the pharynx. This is significant in this research as the drug of interest has an association to this anatomy.

**Table 1: Taxonomy of *Haemonchus contortus*.**

<table>
<thead>
<tr>
<th>Taxa</th>
<th>Name and Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kingdom</td>
<td>Animalia</td>
</tr>
<tr>
<td>Phylum</td>
<td>Nematoda (Diesing, 1861)</td>
</tr>
<tr>
<td>Class</td>
<td>Chromadorea (Inglis,1983) (^{(3)})</td>
</tr>
<tr>
<td>Subclass</td>
<td>Chromadoria (Pearse, 1942)</td>
</tr>
<tr>
<td>Order</td>
<td>Rhabditida (Chitwood, 1933) (^{(3)})</td>
</tr>
<tr>
<td>Infraorder</td>
<td>Rhabditomorpha (Oerley, 1880) (De Ley &amp; Blaxter, 2002) (^{(3)})</td>
</tr>
<tr>
<td>Suborder</td>
<td>Rhabditina (Chitwood, 1933) (^{(3)})</td>
</tr>
<tr>
<td>Superfamily</td>
<td>Strongyloidea (Weinland, 1858) (^{(3)})</td>
</tr>
<tr>
<td>Family</td>
<td>Trichostrongyliidae (Leiper, 1912) (^{(3)})</td>
</tr>
<tr>
<td>Subfamily</td>
<td>Haemonchinae (Skrijabin &amp; Schulz (or Schul'ts or Schul'tz), 1952) (^{(4)})</td>
</tr>
<tr>
<td>Genus</td>
<td><em>Haemonchus</em> (Cobb, 1898)</td>
</tr>
<tr>
<td>Species</td>
<td><em>H. contortus</em> (Rudolphi, 1803) (Cobb, 1898 (type for genus))</td>
</tr>
</tbody>
</table>

The next taxonomic level of the organism in this study as summarised in Table 1 is the Chromadorea (Inglis,1983) class, which usually have elaborate and spiral amphids. The significance is that the anthelmintic under consideration is associated with a physical effect on the amphids. The sophistication in Chromadorea is a factor to consider in the sensitivity and resistance of the organisms to the action of the anthelmintic at and through this organ.

The pharynx is more sophisticated in Chromadorea than other nematodes. The sensitivity of this organ to the action of the anthelmintic being investigated and the potential of a decreased sensitivity and resistance at this location is of interest to this study.

The Rhabditida (Chitwood, 1933) order contains both free-living and parasitic species. In some cases both modes occur in the same species at different life stages such as the free-living second larval stage (L2) feeding on bacteria in faeces and the parasitic adult stage (L5) feeding on the blood of the host in the
studied organism. The beginning of the third larval stage (L3) in the parasitic taxa is infective to the final host, this stage actively climbs grass leaves to be ingested in the pathogenic worm in this work. The Rhabditida contains both the free-living model organism and parasite in this research. The amphids appear pore-like with small openings, this precise anatomy is relevant with regards to changes in this structure associated with anthelmintic mode of action and resistance.

The Trichostrongylidae (Leiper, 1912) family are parasites of herbivores and primarily of ruminants, and as such contain most of the parasites economically important to agriculture. The parasitic species in this work infects sheep and goats alongside other incidental hosts. Trichostrongylidae adults invade the abomasum (fourth stomach) or intestine, the final larval stages and adults of this parasite only occupy the abomasum. The eggs pass through the gut and as Trichostrongylidae juveniles cannot penetrate skin they must remain on the pasture until ingested hence completing and defining the life cycle. The variable proportion of the total population in the host and therefore exposed to anthelmintic compared to the remainder in the environment is relevant to the pressure for selection of resistance to the anthelmintic.

The Strongylus genus was erected with Strongylus contortus (Rudolphi, 1803) but then altered to Haemonchus contortus (Cobb, 1898) as the type for the new Haemonchus genus. The number and synonymy of members has been in flux in the literature with 30 papers raising an additional 30 species, four species inquirendae, two varieties and five subspecies. The majority of these have been synonymised or removed later. The most recent work concludes the Haemonchus genus comprises 12 species with no species inquirendae, varieties or subspecies as detailed in Table 2. However, the species Haemonchus santomei (Gutteres, 1949) has not been dismissed nor investigated since first raised.
Table 2: Haemonchus Species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>H. contortus</em></td>
<td>(Rudulphi, 1803) (Cobb, 1898 (type for genus))</td>
</tr>
<tr>
<td><em>H. longistipes</em></td>
<td>(Railliet &amp; Henry, 1909)</td>
</tr>
<tr>
<td><em>H. placei</em></td>
<td>(Place, 1893)</td>
</tr>
<tr>
<td><em>H. similis</em></td>
<td>(Travassos, 1914)</td>
</tr>
<tr>
<td><em>H. mitchelli</em></td>
<td>(Le Roux, 1929)</td>
</tr>
<tr>
<td><em>H. bedfordi</em></td>
<td>(Le Roux, 1929)</td>
</tr>
<tr>
<td><em>H. vegliai</em></td>
<td>(Le Roux, 1929)</td>
</tr>
<tr>
<td><em>H. lawrencei</em></td>
<td>(Sandground, 1933)</td>
</tr>
<tr>
<td><em>H. okapiae</em></td>
<td>(van den Berghe, 1937)</td>
</tr>
<tr>
<td><em>H. santomei</em></td>
<td>(Gutteres, 1949) not dismissed nor included since</td>
</tr>
<tr>
<td><em>H. krugeri</em></td>
<td>(Ortlepp, 1964)</td>
</tr>
<tr>
<td><em>H. dinniki</em></td>
<td>(Sach, Gibbons &amp; Lweno, 1973)</td>
</tr>
<tr>
<td><em>H. horaki</em></td>
<td>(Lichtenfels, Pilitt, Gibbons &amp; Boomker, 2001)</td>
</tr>
<tr>
<td>Total</td>
<td>12/13</td>
</tr>
</tbody>
</table>

Members of the *Haemonchus* genus are all blood-consuming abomasal (true stomach) parasites of a diverse range of 46 genera of hosts including sheep, goats, cattle, deer, antelopes, camels and giraffes. The type species of *H. contortus* infects the vast majority of these, a total of 40 host taxa.

One of the most globally prevalent and economically important parasitic species in domestic animals is *H. contortus*. In sheep this parasite is the most important as feeding on blood in the abomasum can lead to symptoms of anaemia, lethargy, weight loss and even death in acute infections (haemonchosis).

### 1.2 Anthelmintics

Treatment and control of parasitic worms (helminths) is dependent on de-worming drugs (anthelmintics) as there are no other practical choices,\(^{(5, 6)}\) to the extent that anthelmintics are the most valuable veterinary products constituting 40% of all livestock health expenditure.\(^{(7)}\) One of the three classic anthelmintic classes is the macrocyclic lactones (ML) group that comprises the avermectins (AV) and milbemycins (MI). AVs are 16-membered MLs isolated in the 1970s from soil-dwelling *Streptomyces*. The eight natural AVs demonstrate broad-spectrum activity against ectoparasites and nematodes with avermectin B1\(_a\) being the most potent.\(^{(8)}\) Commercial introduction was in 1981 with ivermectin (IVM) as a mix of two synthetic chemicals, the majority being 22,23-dihydroavermectin B1\(_a\).\(^{(9)}\) The market for IVM has now grown to become the most commonly used drug in
agriculture, with a global consumption of US $1 220 million per annum.\(^7;^9\) It is used for both preventative treatment (prophylaxis) and clinical treatment of ectoparasite infestation and endoparasite infection.\(^7\)

### 1.3 Anthelmintic Resistance

Drug resistance in veterinary helminths is a worldwide and serious problem in agriculture as chemotherapy is essential for agricultural income and animal welfare in the absence of viable alternatives.\(^5\) In particular the widespread resistance to the older anthelmintics including narrow-spectrum drugs and all three broad-spectrum classes is serious in sheep\(^10;^11;^12\) and has increased the reliance on IVM. The emergence of IVM resistance in *H. contortus* is therefore a serious threat to the control of the pathogen\(^11;^13;^14\) to the extent of threatening the entire sheep industry in Australia.\(^13\)

The purpose of chemotherapy is for a drug to be presented to a target to cause an effect. The target and effect can be defined at any level ranging from molecule to organism to population. If an effect does not occur it could be due to the drug, the presentation or the target, and these need to be examined before resistance is declared erroneously. Differences or problems in the formulation, chemical stability or expiration of the drug are the first confounding factors to rule out. Secondly, differences or problems in the presentation including dosage, method of administration and pharmacokinetics (which varies between different host species) need to be eliminated. As parasites are not usually identified before applying anthelmintics, species or life stages may be present and survive as they are not targets for the drug. This example illustrates the fundamental difference between susceptibility and sensitivity. Susceptibility is the first term to be considered as sensitivity only applies in one of the two susceptibility scenarios. If, with no previous exposure to the drug, the drug cannot induce an effect from the target, the target is *insusceptible* only, this must be considered to avoid incorrect and misleading conclusions of resistance. An *insusceptible* target is innately not vulnerable to the drug due to factors present before any use of the drug including protection by the host, the life stage, behaviour, target concentration and distribution, drug metabolism and excretion. A target that is vulnerable to a drug is defined as *susceptible*. A *susceptible* target is innately vulnerable to the drug in at least one host, life stage and season but might be *insusceptible* in others. A
susceptible target is also sensitive if presently affected by the drug. A susceptible target has decreased sensitivity if, on exposure to the drug, acquired characteristics mean the effect has lessened or the same effect requires a higher concentration within therapeutic limits. Once a susceptible target has acquired decreased sensitivity to the point where a drug is ineffective within therapeutic doses the target is finally resistant.

The in vivo test forms the phenotypic definition of resistance, such that the reduction in the number of nematode eggs counted in the faeces following anthelmintic treatment is less than 95% relative to the pre-treatment count (Faecal Egg Count Reduction Test; FECRT). At the present time resistance can be identified and measured on the basis of changes in phenotype such as decreased inhibition of hatching or feeding, or an increase in survival following exposure of *H. contortus* to IVM in vitro. These phenotypic tests are neither sensitive nor reliable enough to detect early selection of resistance but are restricted to being indicators of clinical resistance.

1.4 Genetics of Anthelmintic Resistance

Parasites carrying mutations that decrease drug efficacy in that individual will provide a survival advantage in the face of anthelmintic treatment. This selection for differences in the genetic profile is the first step in drug resistance. The surviving male and female adults are able to reproduce and the genetic profiles are inherited by their offspring. In the case of allelic variation causing resistance, whether the alleles are recessive or dominant will affect whether heterozygotes or homozygotes progeny will also survive. Another control on the genotype frequencies is that in some cases homozygotes are lethal whereas the heterozygotes confer the survival advantage. If anthelmintic is then absent the changes are retained in this heritable shift, so creating a ratchet effect on resistance with an increase, retention then further increase. With selection pressure continuing to be applied, the frequency of the resistant genotypes increases until a significant proportion of the population are resistant to therapeutically relevant doses of drug. This boundary in drug resistance is pertinent with parasites as the dosage has to be below toxic levels for the host. At this stage treatment has failed and the population of parasites in that location is classed as resistant to that anthelmintic.
Sequence variation in parasitic nematodes is high with evidence originally from allozyme analysis and then confirmed by restriction enzyme analysis, single-strand chain polymorphism and now direct sequencing of both mitochondrial and nuclear DNA.\(^{(27; \ 28; \ 29; \ 30; \ 31; \ 32; \ 33; \ 34; \ 35)}\) The diversity at the nucleotide level is the standard measure for reporting\(^{(36)}\) and is at a rate of 0.011 in *C. elegans* but as frequent as 0.026 in parasitic nematodes.\(^{(37)}\) Base diversity is directly related to mutation rate and population size. A mammalian host such as sheep can be infected with thousands of adult trichostrongyloid adult nematodes and one female can pass thousands of eggs per day.\(^{(38)}\) Whilst the effective population size from a genetics standpoint is less than the total number of eggs produced, it is still extremely large in trichostrongylids compared to most other dioecious organisms.\(^{(21)}\) Within the parasitic nematodes the group with the greatest variability are the trichostrongylids, including *Haemonchus*.\(^{(39)}\)

Recombination is a further source of genetic variation correlated with physical distance between loci on the same chromosome, recombination rate per chromosome per meiotic event and the number of meiotic events, which is the number of generations and hence time divided by average life cycle period. Recombination rates are not known for parasitic nematodes.\(^{(21)}\) Linkage disequilibrium is rapidly broken down by recombination in a large effective population and so linkage disequilibrium is inversely proportional to population size. Therefore in the abundant and dense population of trichostrongyloid nematodes such as *H. contortus*, linkage disequilibrium is expected to be short lived.\(^{(21)}\)

Recombination is also a cause of breaking down associations of genetic markers and alleles and thus generating linkage equilibrium. Where particular combinations of alleles are deleterious or beneficial their proportion in the population is less or more than expected by random recombination from the frequency of each allele in the population, creating linkage disequilibrium. In a population under anthelmintic treatment linkage disequilibrium is used as a signal of a genotype conferring resistance. Where a signal does not vary over a physical distance on a chromosome, linkage is present. This is the case in *H. placei*, which has almost complete linkage within 1 kb of a GluCl gene.\(^{(40)}\) Those genetic markers in linkage or linkage disequilibrium in a population exposed to anthelmintic are evidence of selection at or near those loci.\(^{(41; \ 42)}\)
A hard sweep describes a single mutation immediately before or following drug application causing resistance, which is strongly selected for and both the mutation and its associated genetic markers in the original haplotype rapidly rise in frequency in the population. Insufficient time has elapsed for recombination to break up the original haplotype and so linkage disequilibrium and a profound reduction in polymorphism extending out for a large segment from the resistance locus is detected.\(^{(21; 42)}\)

A hard sweep has occurred in *Drosophila melanogaster*,\(^{(43)}\) *Plasmodium falciparum*,\(^{(44; 45)}\) rats\(^{(46)}\) and in both the trichostrongylids *Teladorsagia circumcincta* and *H. contortus* a predominant haplotype for BZ resistance has been identified.\(^{(47)}\)

A soft sweep is the selection of multiple different haplotypes conferring drug resistance and seems more likely than a hard sweep in anthelmintic resistance.\(^{(21; 42; 48; 49)}\) Multiple haplotypes comprise a complex scenario where the genetic footprint is weak and hard to detect.\(^{(21)}\) However, whilst the reduction in polymorphism is minimal, an unusually strong pattern of linkage disequilibrium is discernable.\(^{(49)}\)

The origin or multiple origins of drug resistance could have been in existence prior to the use of anthelmintics, a novel change just prior to or during treatment, or recurrent mutations. In addition migration can introduce any of these three origins onto a farm or geographic area through movement of the livestock hosts. The origin of anthelmintic resistance is important as it affects which methods are most efficient in investigation, surveillance, control and treatment. The candidate gene approach would best suit a single genetic origin whereas genome-wide studies would assist with multiple or recurrent causes. Novel or recurrent origins would dictate that on-farm control is more important than quarantine measures, whereas pre-existing alleles that can be introduced by livestock are best protected against by biosecurity and treatment followed by testing for any remaining nematodes during quarantine.\(^{(21)}\)

The case for a recent spontaneous origin of anthelmintic resistance is the plentiful and continual supply of mutations in nematodes with a large population and high mutation rate.\(^{(21)}\) The mutation rate in nematodes is higher than other
phyla\(^{(30; 39; 50)}\) in both mitochondrial and nuclear DNA in \textit{C. elegans}\(^{(51; 52)}\). The polymorphism rate per base pair has been measured as \(9 \times 10^{-9}\) in \textit{C. elegans} and calculated as 2 base pair substitutions per nuclear genome per generation\(^{(53)}\) or one SNP for every base on average in \(5 \times 10^{-7}\) individuals. A sheep with a moderately high \textit{H. contortus} burden would shed that number of eggs in a few days demonstrating the variability in the population. Data from the \textit{H. contortus} genome sequencing project reveals insertions and deletions (indels) from a few bases to many kb are commonplace. The mutation rates for the model organism are underestimated as half of mutations are indels in addition to SNPs, therefore the potential for both single and multiple independent resistant alleles is evident.\(^(21)\)

A direct demonstration of multiple independent origins of resistance in \textit{T. circumcincta} and \textit{H. contortus} is the presence of two different resistant alleles on one goat farm and alleles unique to different goat farms closed to livestock movement.\(^{(47)}\) The case for pre-existing alleles lost by genetic drift in other farms is unlikely due to the parasite population size and BZ selection pressure on those other locations.\(^{(21)}\) Current mathematical modelling also supports multiple independent origins especially with the large population size and high mutation rate in trichostrongylid nematodes.\(^{(48)}\) More evidence for independent origins of BZ resistance is the difference in \(\beta\)-tubulin isotype 1 restriction fragment patterns between geographically separate \textit{H. contortus}.\(^{(22)}\)

The case for pre-existing alleles is strengthened by the ease and speed of selection for IVM-resistant \textit{H. contortus} in just three generations.\(^{(54)}\) The rapid emergence of tyrosine at codon 200 in \(\beta\)-tubulin isotype 1 preventing BZ binding and conferring resistance in \textit{H. contortus} is most likely explained by pre-existing alleles at relatively high frequencies.\(^{(21; 23)}\) The separate resistant alleles on the goat farms that suggest recurrent emergence are also consistent with pre-existing genotypes.\(^{(21; 47)}\) The other argument from that work was the farms being closed to hosts meant the mutations had occurred independently is also explained by the sequence being present prior to closure.\(^{(21; 47; 55)}\) The populations of \textit{T. circumcincta} had little genetic differentiation between the farms, thus supporting the hypothesis that the same genetic origin was being registered.\(^{(21; 34; 56)}\)
Migration of resistance has been proved with a single allele being discovered worldwide in insects.\(^{57}\) With the tremendous amount of movement of livestock around the UK, the seven different resistant haplotypes in \textit{T. circumcincta} in open UK flocks are postulated to be due to migration via ruminants.\(^{21}\)

The number of stages and processes outlined belie the evidence that anthelmintic use inevitably and rapidly leads to resistance.\(^{54; 58; 59}\) Anthelmintic resistance is most advanced in \textit{H. contortus} and the understanding of the phenomenon in parasites is greatest for this species.\(^{58; 59}\)

\textbf{Table 3: Evidence of the Genetics of IVM Resistance.}

<table>
<thead>
<tr>
<th>Drug</th>
<th>Genetic Evidence</th>
<th>Species</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>IVM</td>
<td>ligand for P-gp</td>
<td>in mice</td>
<td>((60; 61))</td>
</tr>
<tr>
<td>IVM</td>
<td>inhibit transport function of some P-gp</td>
<td>in mice</td>
<td>((60; 61))</td>
</tr>
<tr>
<td>IVM &amp; MOX</td>
<td>associated through population genetic studies with P-gp</td>
<td>in \textit{C. elegans} &amp; \textit{H. contortus}</td>
<td>((62; 31))</td>
</tr>
<tr>
<td>IVM</td>
<td>not consistent association through population genetic studies with P-gp</td>
<td>in \textit{H. contortus}</td>
<td>((63))</td>
</tr>
<tr>
<td>IVM</td>
<td>selection on ABC transporter</td>
<td>in \textit{O. volvulus}</td>
<td>((64, 65, 66, 67))</td>
</tr>
<tr>
<td>IVM</td>
<td>decreased polymorphism of P-gp</td>
<td>in \textit{H. contortus}</td>
<td>((31; 68))</td>
</tr>
<tr>
<td>IVM</td>
<td>decreased polymorphism of P-gp</td>
<td>in \textit{O. volvulus}</td>
<td>((69))</td>
</tr>
<tr>
<td>IVM</td>
<td>overexpression of P-gp</td>
<td>in \textit{H. contortus}</td>
<td>((70))</td>
</tr>
</tbody>
</table>

The genetic basis of IVM resistance in \textit{H. contortus} is not currently known,\(^{6; 71}\) which precludes any monitoring of resistant genotypes prior to phenotypic changes. The partial evidence of the genetics of IVM resistance is summarised in Table 3. One candidate are the P-glycoproteins (P-gp), these are members of the adenosine triphosphate (ATP) binding cassette (ABC) transporter superfamily and as such are efflux pumps in cell membranes, removing hydrophobic xenobiotics from the cell membrane and cytoplasm.\(^{72; 73}\) There has been overwhelming evidence of their rôle in anthelmintic activity, IVM is a potent ligand and inhibits the transport function of some P-gps.\(^{60; 61}\) P-gps have been associated with ML resistance from population genetic studies\(^{31; 62}\) although this was not consistent.\(^{63}\) However, in the nematode parasite \textit{Oncophora volvulus} selection of IVM on an ABC transporter was verified\(^{64; 65; 66; 67}\) and in both \textit{O. volvulus} and
H. contortus IVM treatment led to decreased polymorphism of P-gps, a classic indication of selection.\(^{31;68;69}\)

### 1.5 Study of Genetics of Anthelmintic Resistance

There are a number of approaches to studying the genetics and genomics of anthelmintic resistance as summarised in Table 4 and detailed below.

**Table 4: Approaches to Studying Genetics and Genomics.**

<table>
<thead>
<tr>
<th>Option</th>
<th>Alternatives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Relationship</td>
<td>Inheritance Linkage</td>
</tr>
<tr>
<td></td>
<td>Association</td>
</tr>
<tr>
<td>Region</td>
<td>Candidate</td>
</tr>
<tr>
<td></td>
<td>Genome wide</td>
</tr>
<tr>
<td>Variable</td>
<td>Allele</td>
</tr>
<tr>
<td></td>
<td>Expression</td>
</tr>
<tr>
<td>Discrete or Continuous</td>
<td>Case-Control</td>
</tr>
<tr>
<td></td>
<td>Quantitative</td>
</tr>
<tr>
<td>Outcome</td>
<td>Final</td>
</tr>
<tr>
<td></td>
<td>Intermediate</td>
</tr>
<tr>
<td>Trait</td>
<td>Phenotype</td>
</tr>
<tr>
<td></td>
<td>Genotype</td>
</tr>
</tbody>
</table>

An inheritance study investigates the genetic linkage to phenotype (not physical linkage on the chromosome between genes) through the comparison of parental genetics and their traits to that of the offspring. Classically, the link between candidate genes and outcome was established in this manner. However, genome-wide comparisons between parents and progeny have been made and could be for *C. elegans* in the lab and with sheep artificially infected with one male and one female *H. contortus*. This prevents use for investigating resistance in field populations under normal farming practice. Populations of known lineage with sensitive parental populations and resistant progeny in as little as three generations separation have been specifically retained as sources for inheritance linkage studies. The genetic linkage studies succeeded with single gene traits but were less reliable for multiple and complex causes of a phenotype.

Due to the limitations of genetic linkage studies, association between genotype and trait was initiated. As signals were expected from multiple genes combining in a complex interaction to produce an outcome the study of association was required. The populations are differentiated on trait with no knowledge of relationship or inheritance. The genotypes and traits of each individual are compared and the number of instances of each of the combinations of each genotype and traits established. The odds ratio of a genotype being present with a trait is calculated and the significant pairings are reported as associated.
Unfortunately, the vast majority of associations have low odds ratios, such that the genotype has a low effect on the trait being measured. This is especially the case when common polymorphisms are being investigated. Even in traits with known heritable components the association genes contribute relatively little. With a potentially limited contribution to the trait such as anthelmintic resistance, a limited contribution of genotype reduces the utility of such a finding as a molecular test of IVM resistance for example. As such the sensitivity is expected to be higher with fewer false negatives. In contrast the number of associations is more frequent and if association is carried out on a whole genome the number of false positives is increased and the specificity reduced. A limitation of association is that causality cannot be inferred and hence further more traditional and focussed molecular genetic work is required. However, it has a use as a first pass, particularly across the whole genome. Common polymorphisms are usually the subject matter in association studies and whilst these common alleles may be contributing, a more rare allele could not be included yet have an important contribution. Due to the restrictions of association, particularly of common SNPs, large numbers of organisms are required to provide sufficient sample sizes. With the population density of *H. contortus* this is not difficult but good definitions of cases and controls or of a quantitative trait are required.

Both association and inheritance linkage genotypes can be in the gene influencing the trait but it could also be close on the chromosome to the gene with a rôle, these two would be in physical linkage. Alternatively, the genotype could be in a non-random linkage with a locus not physically adjacent, if despite meiotic recombination the pattern remains and the two would thus be in linkage disequilibrium.

Populations should be controlled for sex, life stage and for population stratification from differing geographic sources. The genetic structure of the *H. contortus* global population is reported as a complex of biotypes, relevant to the population stratification that can be a confounding factor in association research.

A candidate gene approach uses existing knowledge to choose genes to specifically investigate as opposed to a whole genome approach. The knowledge can be sourced from a priori knowledge of the direct impact of a gene or gene family on a trait such as β-tubulin isotype 1 mutation with BZ binding and
resistance. This avenue can be limited so a more indirect approach is utilising known biological, physiological, functional or pathophysiological rôles of the genes that are relevant to the trait in question such as the IVM being a good substrate for the P-gp pumps. The sources may well be inconclusive but still point to genes of interest.

The obvious restriction to a gene being a candidate is reliance on previous knowledge and understanding of all the genetic contributions to a trait. Where multiple or complex interactions cause the trait an individual gene may not provide a clear association of inherited linkage. This can be counteracted by assessing several genes or families of genes simultaneously. Candidate gene studies often investigate sequence variation but can and do also measure expression.

Genome-wide or whole-genome studies consist of three main forms, studies of common variation, complete genome sequencing and microarrays. The majority of common variation is SNPs but such analysis requires knowledge of a substantial number of alleles and loci across the entire genome. By definition, sites of common variation will generate a large dataset across a genome and this sensitivity can provide opportunity for false positives. This method can be applied to association or inheritance linkage processes. In the case of association the number of results is impressive but the odds ratios can be less so.

One use for genome-wide sites of polymorphism is as a discovery cohort with no bias from pre-existing knowledge or assumptions. The alleles with the greatest and most significant odds ratios are then investigated further in a validation cohort. The use of a non-hypothesis-driven method is attractive in the case of IVM resistance as the progress in the gene families thus far has been limited.

With technological developments the alternative to partial sequencing is whole genome sequencing in high-throughput systems. The analysis of such data sets is an even more critical step. The first full genome is yet to be completed for H. contortus but high-throughput approaches can be applied to unknown genomes and the sequences then analysed on a project by project basis.
Microarrays detect changes in expression of transcripts genome-wide. The sensitivity to differences, number of results and the resulting potential for false positives are areas of concern.

Instead of a discrete variable being compared such as IVM resistance or sensitivity, a quantitative trait can be measured such as the concentration of IVM to cause an outcome or the measured motility in a set IVM concentration. The benefit of such an approach would be detection of intermediate populations with a decreased sensitivity but not resistant within therapeutic doses. This is useful in the study of hypotheses to control or prevent resistance and in surveillance where sub-clinical problems being identified could inform farmers and vets of the need to act. Where loci are being mapped with a quantitative trait the process is termed quantitative trait locus mapping, QTL mapping.

The trait being considered is often a phenotype such as the number of eggs, motility or development from one life stage to another. However, the outcome to compare can be a genotype. In this case variation in sequence or expression can be investigated with the expression of a particular transcript as the outcome. The process of mapping loci with the quantitative trait of the expression of a particular gene is termed expression quantitative trait locus mapping, eQTL mapping.

There are opportunities for non-candidate results to feed into candidate-driven work. Non-biased suggestions are then investigated more thoroughly. Genome-wide association studies can indicate associations between the trait and several possibilities. These new regions of interest can then be candidate loci and genes. The initial signal can be within a gene with potential influence or be in physical linkage with another gene. This physical linkage can be hypothesised when a genome provides whole chromosomai sequences, which is still to be completed for *H. contortus*. A linkage disequilibrium could also be the indirect connection and would need further investigation to decide. Quantitative trait locus mapping and microarrays can also distinguish sites as candidates for a causative role or being in physical linkage or linkage disequilibrium with a relevant gene or gene complex.

An expansion to the concept of the candidate gene being based on previous work is to be thorough in the bioinformatic analysis of the sequences metadata,
annotation and reported results across the genome, transcripts, genes and gene families to detect signals for candidate genes.

There are multiple areas requiring attention for the better prevention, detection and control of anthelmintic resistance. Tests need to be developed that are both reliable and practical in terms of equipment, time and cost in order to be used widely rather than when resistance is already established. The genetic mechanisms should be elucidated so that molecular tests can be developed that are sensitive to population changes prior to clinical manifestation. More generally an improvement is sought in the knowledge and understanding of the biology, ecology and epidemiology of parasites that could then allow for better interventions and provide conclusive answers to some of the debates surrounding parasite control. Finally, encouraging and creating systems and communication channels of expertise would allow these new lessons to be disseminated.\(^{6}\)

\[1.6\] Aim of Project

As described above, one of the areas requiring attention for the better prevention, detection and control of anthelmintic resistance is the elucidation of the genetic mechanisms so that molecular tests can be developed that are sensitive to population changes prior to clinical manifestation. In the absence of genomic tools currently available for determining genes under IVM selection in \textit{H. contortus}, the approach taken to date is the analysis of the individual genes implicated in IVM resistance including the P-gp family. \textit{H. contortus} has an incomplete genome without annotation and only one P-gp mRNA transcript has been sequenced.\(^ {70}\) The aim of this project was to discover the incomplete and dispersed P-gp genetic sequences in the sequence databases and construct the putative genes. These could then be investigated in the future for their rôle in IVM resistance in \textit{H. contortus} and potentially utilised in molecular tests for IVM resistance.
2 MATERIALS AND METHODS

2.1 Bioinformatics

The purpose of bioinformatics was to discover novel genetic sequences in the incomplete *H. contortus* genome that map both from and to known P-gp sequences and therefore form putative *H. contortus* P-gp genes (*pgp*).

*C. elegans* and *H. contortus* are both in phylogenetic clade V of the nematodes and *C. elegans* is commonly used as a model organism for *H. contortus*. The genome of *C. elegans* having been completed and annotated allowed extraction of the entire annotated *pgp* gene family from *C. elegans*, which was used to expand the P-gp protein (PGP) family in *H. contortus* from the single member (previously termed *Hco*-PGP-A) based on messenger RNA (mRNA).

Multiple genomic DNA (gDNA) databases existed for *H. contortus* as the nucleotide sequences were generated using six different methods and batches of new sequences released at seven different times between December 2004 and August 2009. There were a total of 21 separate searchable databases, comprising 12 distinct types of sequences with the remaining 9 databases having been included in and superseded by more recent collections.

2.1.1 Alignments along Entire Gene

The amino acid sequences for the PGP family in *C. elegans* of 14 proteins and the translation of the single pseudo-gene (*Cel-pgp-15ps*) were retrieved from the WormBase database (www.wormbase.org). These sequences were used to trawl the incomplete *H. contortus* genome databases hosted at the Wellcome Trust Sanger Institute (Sanger) for similar sequences using the Washington University-Basic Local Alignment Search Tool (WU-BLAST) (http://www.sanger.ac.uk/cgi-bin/blast/submitblast/h_contortus). The specific WU-BLAST algorithm used was TBLASTN, in order to search translated nucleotide databases with the protein queries. (TBLASTN translated the nucleotide databases into all six reading frames and then used these amino acid sequences as the subject for a search using the amino acid sequence entered by the author.) The matrix used for amino acid substitution scores was the standard BLOSUM62.
The biological interest in an alignment was whether it provided evidence of homology between the underlying genes, however, the question has to be inverted, as is common in statistics, into the question of how much evidence existed that the alignment was due to chance alone. Continuing that logic meant that an alignment expected to occur less frequently by chance in comparison to another alignment was more likely to be homologous and biologically meaningful. To report that homology and significance, each alignment had an Expected value (E-value) calculated as below.

The query sequences are the amino acid identities of the known *C. elegans* and *H. contortus* proteins plus the translated pseudo-gene *Cel-pgp-15ps*. One amino acid sequence is compared to the subject database of *H. contortus* nucleotide sequences translated into all six reading frames. The amino acid residues are compared with a particular scoring matrix, BLOSUM62. Identical residues receive a highly positive integer score, residues with similar physico-chemical properties receive a moderately positive integer score and pairs of residues with differing properties are scored negatively. The BLOSUM62 scores were originally chosen to provide a negative sum for lengths of residues with no structural or functional similarity and thus a random alignment provides a negative total and is discarded. A word length of 3 was used to seed alignments, for those that exceed a minimum positive score the alignment is tested whether extending in one or both directions increases the total score, this can involve a shorter length of fewer negative pairings in which case the alignment is trimmed to obtain the maximal score (S).

\[ S \]

The resulting alignments that passes a minimum total score threshold are retained as the high scoring segment pairs (HSPs), the remainder are discarded. Each scoring matrix has a particular statistical effect on how the resulting scores compare to the minimum amount of information that is required to convey a range of alignments. This value for BLOSUM matrices is calculated in advance on random alignments and has a set value for alignments on a particular database. This number enables different scoring systems to be compared and is termed \( \lambda \).
\[ \lambda \cdot S \]

The total score for an alignment multiplied by \textit{lambda} is the exponent for calculating the number of alignments to be expected from chance alone.

\[ \exp(\lambda \cdot S) \]

The exponent is negative such that a large positive score generates a small number of alignments expected from chance, a doubling in the score reduces the expected number. The expected value therefore approaches zero and frequently is reported as zero but is in fact always at least an infinitesimally small positive value.

\[ \exp(-\lambda \cdot S) \]

The exponential function makes sense as for a score to double is the same as half the score occurring twice on the same alignment. The occurrence of that scenario due to chance alone is exponentially less likely than the best alignment only having available the score once.

The length of the subject database plays a role in the chance of an alignment as a larger set of sequences or a longer average sequence increases the chances in a linear function of an alignment with a certain score or above. A doubling of the total search space doubles the number of alignments with a certain score due to chance and hence doubles the expected number of alignments. For protein sequences the length of sequences in the database is limited to the range of protein lengths and is not arbitrary. The assumption in BLAST is that the lengths of the component individual sequences within the massed total length \( (n) \) is relevant.\(^{77} \)

\[ n \cdot \exp(-\lambda \cdot S) \]

\[ n^{(-\lambda \cdot S)} \]

However, for genomic DNA databases the individual sequence lengths are arbitrary. The only practical calculation for calculating the number of expected
alignments to a DNA subject database is that the total massed sequence length is relevant and the individual sequence lengths are irrelevant. This holds true for very long gDNA lengths where the individual sequences lengths approach the maximal quantities. It is worth noting that for this research utilising on-going data releases of DNA databases the sequence lengths changed radically from thousands of bases in the contigs of 2006 (CON) to the hundreds of thousands of bases in the supercontigs of 2009 (S09). The calculated expected number of alignments was increasingly accurate.

The high scores are less likely near the ends of a sequence as the missing alignment reduces the positive score. This edge effect is minute in full genome coverage but increasingly pronounced with the more disjointed earlier databases.\(^\text{(76)}\)

The characteristic of a database in producing expected numbers of alignments from chance is captured by the parameter \(K\). Multiplying this with the expected number enables the appropriate scaling of the expected number and allows comparison between different databases.

\[ K \cdot n^{(\lambda S)} \]

The length of the query sequence \((m)\) is clearly relevant to how many alignments would be expected from chance. Increasing the query length provides more lengths of sequence that could align with a particular score or higher. This relationship is such that doubling the length doubles the expected number of alignments.

\[ K \cdot m \cdot n^{(\lambda S)} \]

This function now generates the expect value (E-value) of the number of alignments with the particular score \((S)\) or higher, taking into account the characteristics of the scoring system \((\lambda)\) and subject database \((K)\), the length of the query sequence \((m)\) and the total massed length of the subject database \((n)\).\(^\text{(78)}\)
E-value = K·m·n^{-(λ·S)}

**Equation 1: E-value for Alignments.**

Those seeded words exceeding the minimum initial score, then extended to obtain a maximal score locally, then exceeding the minimum score are the HSPS and those HSPS with an expect value exceeding the threshold are reported in the BLAST results.

The E-value (incorporating a capital E) was provided in shortened exponential notation (incorporating a lower case e to represent exponent) e.g. 2e-3 equals 2 times 10 raised to the power of -3. This E-value was used to rank the alignments and thus the degree of homology between genes from other species and the *H. contortus* sequences: the closer the E-value was to zero the greater the homology inferred. The *H. contortus* sequences that had greater homology in the *back BLAST* with genes outside the *pgp* family were discounted in the search for the *H. contortus pgp* gene family, this removal of noise is the benefit of a *reciprocal BLAST*. Those sequences that matched translated *pgp* genes better than other gene families were putatively identified as components of the *H. contortus pgp* genes.

### 2.1.1.1 Alignments from Assembled Contigs and Expressed Sequence Tags Databases

In this first phase of bioinformatics the two sources most likely to reveal substantial information were investigated; these were the *assembled contigs* database (contiguous sequences derived from shotgun sequencing; released on 27 January 2006 and the most recent at the time; CON) and the expressed sequence tags (EST) database. The *contigs* database contained short fragments of 1-3 kilobase pairs (kb) length and the ESTs were typically only several hundred base pairs long. The results were returned with amino acid sequences from *H. contortus* (as translated from nucleotide sequences) aligned with amino acid sequences from *C. elegans* (as entered from proteins or the translated pseudo-gene). The alignments generated by WU–BLAST between *C. elegans* and *H. contortus* from the Sanger databases contained both relevant and irrelevant matches. To test whether a match was from a relevant putative gene, a second search in the opposite direction was utilised and termed a *back BLAST*, the combined searches
were described as a \textit{reciprocal BLAST}. The \textit{H. contortus} sequences were manually searched for highly similar or identical regions as the aim at this stage was to discover short identifying fragments for every gene rather than the entire gene sequences. These short amino acid sequences were provided as the query for \textit{back BLAST} searches against the nucleotide collection database without restriction to any species; the program and database were hosted at the National Center for Biotechnology Information (NCBI; blast.ncbi.nlm.nih.gov/Blast.cgi). In order to ignore the expected differences in DNA for the same amino acid sequence between the species the \textit{H. contortus} translated nucleotides were used as the query and a translated nucleotide database were used as the subject for these searches, which therefore required the TBLASTN algorithm. As the two databases investigated initially did not return homologous sequences long enough to be useful the other databases were investigated.

\textbf{2.1.1.2 Alignments from Eight Databases}

The next phase of bioinformatics compared the maximum length of alignments returned and coverage of the genome from all eight \textit{H. contortus} databases available at Sanger at the time: the three \textit{sequence reads} databases (released in December 2004, SEQ1; August 2005, SEQ2; November 2005, SEQ3); the \textit{unassembled reads} database (released in January 2006; UNA); the \textit{assembled contigs} database (released in January 2006; CON); the \textit{BAC end sequence reads} database (bacterial artificial chromosome; BAC); the \textit{fosmid end sequence reads} database (FOS) and the \textit{expressed sequence tags} database (EST).

A larger batch of sequences was released as the \textit{assembled contigs all reads} database (ASS) in November 2007. This database was queried with the entire \textit{C. elegans pgp} gene family using TBLASTN. For the \textit{back BLAST} the BLASTX program was run with the nucleotide sequences (translated by BLASTX) from \textit{H. contortus} as the query against the non-redundant protein sequences (\textit{nr}) database without restrictions to species of origin (program and database hosted at the NCBI) and thereby completed the \textit{reciprocal BLAST} method. The E-value of aligned sequences was utilised to assess homology with \textit{C. elegans} PGP proteins. More detailed and sensitive alignments were obtained using the BLASTX version of the BLAST2 program at NCBI, these were manually edited and cropped. An overlap between two sequences could have been insufficient for genome analysis.
to categorise the pair as contiguous. However, the terminal sequences could have been sufficient for more general bioinformatic tools to detect an overlap. Hence both the Contig Assembly Program (CAP3), a component of the Mobyle platform (mobyle.pasteur.fr), and the BLAST2 program at NCBI was run on the entire set of contigs to detect overlaps, even if incorporating single nucleotide polymorphisms (SNP), as a precautionary measure.

2.1.2 Alignments to Specific Domains

Following the BLAST attempts at alignments, an alternative approach was required. Despite the presence of polymorphisms the P-gps, in common with all proteins, contained functional residues that were highly conserved. In addition, whole regions have particular functions and the range of variation at each residue is limited if the function of that domain is to be retained and for there not to be a fitness and survival penalty. This was exploited in the alignment of genomic sequences to known domains.

2.1.2.1 Alignments to Inter-nucleotide Binding Domains

Several sequences were known for the inter-nucleotide binding domains (IBD) in *H. contortus* and were used to solve the gDNA sequences in that region. Prior to use in *H. contortus* the process was evaluated for specificity with comparison of the results between amino acid and nucleotide investigations for *C. elegans*. The *C. elegans* amino acid sequences from each IBD were compared via the BLASTP software (designed for amino acid queries of amino acid subjects) at NCBI with the full-length *C. elegans* PGP protein sequences then the IBD amino acid sequences were aligned by TBLASTN to the translated nucleotide *pgp* sequences. The alignments and E-values were analysed for perfect matches and therefore confirmed the system could be trusted for aligning *H. contortus* sequences. The nucleotide sequences at Sanger were interrogated with the *H. contortus* IBD amino acid sequences by TBLASTN and both the text alignments and the E-value employed to determine which IBD amino acid sequences were represented within the assembled contigs all reads database (ASS). To corroborate the alignment the aligned sequences were assessed in their pairs via BLAST2.
2.1.2.2 Alignments to ATP Binding Cassettes

Further investigation of the possibilities of domain matching as an improvement on simple alignments included using the ScanProsite tools to probe the PROSITE database (provided by ExPASy; www.expasy.org/prosite). The next step was the exploration of the Conserved Domains Database (CDD) by means of the Conserved Domain Search (CD-Search) software at NCBI (www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). A conserved domain (cd) record cd03249 specific to P-gp was then utilised in the next phase. Whilst CD-Search was indispensable as a sensitive detection tool, the fact that cd03249 was a Position-Specific Scoring Matrix (PSSM) and could be depicted using the PSSM Viewer at NCBI (www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm_viewer.cgi) revealed necessary detail of the individual scores and hence how fitting a residue was at each position. Furthermore, the PSSM Viewer permitted alignment of amino acid sequences to the PSSM and was able to generate E-values. This approach was the most apt for the P-gp family as synonymous SNPs were disregarded and non-synonymous SNPs were assessed according to the matrix of scores depending on the specific position of the variant amino acid. This analysis was considered to come closer to a test of function and so more reliable as an indicator of gene identity than the direct study of sequence.

The translated contigs previously extracted from the H. contortus ASS database and putatively aligned with C. elegans were aligned with cd03249 using PSSM Viewer. These multiple alignments were arranged in parallel to both contrast the residues present at the same positions and also to gauge the number of domains present and hence half that number of pgp genes. The Jalview alignment editor was utilised for further manipulation of the C. elegans and H. contortus domains. The homology between the putative and annotated genes was approximated by a phylogeny inferred through computation of individual distance matrices (Jones-Taylor-Thornton model) for each H. contortus sequence with all the C. elegans ABC domains using the protdist component of the Phylogeny Inference Package (PHYLIP) hosted at Mobyle (mobyle.pasteur.fr/cgi-bin/portal.py?form=protdist). The initial alignments were replicated by 1 000 boot-strapping cycles and a distance matrix calculated for each in order to measure the robustness of the neighbour-joining consensus tree that was generated by the neighbor program offered at the same Mobyle site. This evidence of homology was used to group
the contigs according to C. elegans homologue. The disjointed sequences were linked into putative genes by inspection of the adjacent ends for evidence of a clean join but more commonly the intron-exon boundaries required correction and small gaps still persisted. At this juncture the H. contortus sequences from the ASS database composing the ABC had been detected and aligned.

2.1.2.3 Alignments to Both ATP Binding Cassettes and Transmembrane Domains

The method of using the cd03249 PSSM to detect the ABC domains had revealed a portion of the proteins but pairing all the domains had not been possible. The method was therefore repeated with CD-Search employed to reveal other conserved domains specific to P-gp that were then aligned by the PSSM Viewer. A conserved domain (pfam00664) for the ABC transporter transmembrane domain (TMD) and for the entire multidrug resistance protein (a P-gp) in Plasmodium falciparum (PTZ00625) were investigated but a more sensitive matrix was discovered in the ABC-type multidrug transport system model (COG1132) that represented both the ABC and the TMD. This combined matrix meant those H. contortus sequences from the ASS database already aligned with the ABC domain were able to be extended into the TMD and additional sequences were detected that only aligned within the TMD. However, full resolution of the TMD was restricted due to the lower conservation in comparison to the ABC.

This method was repeated with the longer contigs from the assembled contigs database (C09) when it became available in August 2009. With longer sequences the benefit was greater than just the proportional increase in contig length or percentage coverage of the genome. This was because a greater proportion of the longer sequences covered both the ABC and TMD, even if only partially, and therefore the PSSM Viewer was more sensitive and able to increase the coverage of the pgp gene family. However, the C09 database was still too fragmented to provide full-length alignments and gaps remained.

The supercontigs sequences in the S09 collection comprised contigs of known length and nucleotide identities (present in the C09 database), separated by regions of known length but unknown sequence, designated by strings of n. This evidence removed the previous necessity of individually determining which contigs...
were adjacent by manual methods as with the ASS contigs. Due to the length of these sequences (despite the unidentified regions) alignment to full-length genes was again attempted, similar to the original method with the much shorter sequences from the ASS database.

2.1.3 Alignments along Entire Gene

The 16 known PGP amino acid sequences (Hco-PGP-2.1.A, Cel-PGP-1 through 14 and Cel-pgp-15ps) were queried against the supercontigs database (S09; released in August 2009) using WU-TBLASTN, hosted at Sanger. The sequences with E-values below 10 were extracted from the database; although this E-value was high and allows false positives (lower specificity) it was also less likely to cause false negatives and so was more sensitive. Every putative pgp sequence from this forward BLAST was then compared to the C. elegans nr database. Sequences with at least regional matches (higher S scores; lower E-value) to pgp rather than other gene families were considered to contain pgp exons. The sequence was then compared to the amino acid sequences (Hco-PGP-2.1.A, Cel-PGP-1 through 14 and the translation of Cel-pgp-15ps) in order to determine putative homologues. Whilst the reciprocal BLAST methods were more sensitive to these sequences of greater coverage, the need for manual editing of the alignments and determining misalignments, particularly around the intron-exon boundaries and the unidentified regions was still necessary.

2.1.3.1 Gene Structure Prediction Algorithms

The alternatives to BLAST and manual editing included mGene, Splign, Spidey, DiAlign and GeneWise; all were evaluated for use in predicting the pgp gene sequences in H. contortus but only the GeneWise program was productive in this aim.

The purpose of GeneWise was as a comparison of genomic sequence with a homologous protein in order to predict the gene sequence and structure. The GeneWise method integrated two different processes sequentially, the first being a model of alignment of a homologous protein to the translated genomic sequence and the second is a model of predicting the gene structure.
All possible translations of the genomic sequence in all reading frames were considered against the homologous protein in order to generate a predicted amino acid sequence. Accepting all reading frames at every base allowed for insertion or deletion errors to not cause the remainder to be out of phase. These nucleotides were taken account of by a codon being allowed to be 2, 4 or 5 bases where the codons upstream and downstream provide a better alignment. However, no attempt was made to translate the abnormal codon because the calculation of all possibilities is too complicated and computationally expensive. The codon thus provided a null residue. This approach was far more resilient to resolving a genomic sequence to exons that pass over errors.

The alignment model does not attempt to locate substitution errors originating from the genomic sequencing. Instead the probability of a transition or transversion was included within the calculation for a triplet to be aligned with an amino acid. For those triplets where a reported base was different from the organism the effect on the probability assigned to a particular translated sequence and the scoring of that alignment was spread over a wider locus than the actual nucleotide. This smudged effect is of no consequence to the predicted gene structure.

GeneWise did not itself provide the profiles of protein homology for the model to use but instead incorporated an established ‘worm’ pattern. The protein model was affected by three considerations: amino acid bias in the protein provided, codon bias in the organism providing the genomic material and sequencing error in the determination of that genomic sequence. The amino acid bias and codon bias are incorporated into the probabilities being calculated for a translated triplet being aligned with a base. The general ‘worm’ model provided the bias parameters and the default sequencing error rates of $1 \times 10^{-5}$ was used as it is the reported standard for genome sequencing.

The second part of the amalgamated algorithm predicted gene structure. The aligned protein, splice site patterns and parameters for the genomic sequence fed into a maximum likelihood estimate of the various splice sites along the DNA sequence. A related approach from the protein alignment model was applied in the gene prediction model for splice sites in phase 1 and 2 that therefore split a codon by the intron. The signal from both 5’ and 3’ splice sites were received and contributed to the predicted gene structure. The translation of the codon was then
ignored to avoid a further submodel for just these calculations of identity. The position in the alignment over the splice site codon was not scored to protect against a negative score when the amino acid cannot be aligned to a non-diagnosed identity. A negative score would have reduced the evidence for this alignment and the overall score, in the case of marginal alignments this could have swung the evidence to report alternative exons or end the alignment early.

A number of the protein alignment model and gene structure model outputs were not possible, for example nonsensical genetic sequence. These states were forced to zero and then removed. Also some theoretical intermediate or final cases were not required, for example concerns of multiple changes occurring in one codon where the greatest portion of the possibilities have infinitesimally small probabilities. These were also forced to output a probability of zero without calculation and hence the states are not present in the process.

The gene structure algorithm had the ability to distinguish further signals of the central intron, polypyrimidine tract and the spacer at the end of the intron. The sequential application of the homologous alignment and the gene structure models including the intron signals would have resulted in 21 states and 93 transitions in total. That number of calculations was ambitious and slow. More significantly, the intron signals were too sensitive to general gene structure in the genomic sequence provided despite the homology being distant and the alignment poor. Determining the exact position of an indel sequencing error where a 2, 4 or 5 base codon has been utilised was also too expensive. The error locus was likely to be 1 position different from that reported by the model but that was of no consequence. Losing the intron and exact error positioning states and transitions brought the combined model down to a manageable 6 states and 23 transitions. This version is termed GeneWise 6:23 or 623 in the parameters reported in the results and appendix and was the version employed in this study. The functional use of this model has been reported as being excellent in previous studies with the edge detection of exons providing a strong signal principally through the splice sites.

There were several modes available in GeneWise to account for different alignment and gene structure requirements. In the expansive model of 21 states and 93 transitions all genes score well. Mispredictions could therefore occur both in terms of any alignment to non-homologous proteins but also extended coverage
of the genomic sequence with the predicted gene, especially with the terminal exons being predicted for the best start and stop signals and splice sites. A specific flanking model as an addition to the combined model was too expensive. The global mode did allow for stretched genes but was useful as an initial choice to detect the core alignment. The wing approach was to not penalise alignments that start or end within the terminal five bases and hence allowed truncation and downstream starts and upstream stops, this mode was not more effective than other modes and was not utilised. Endbias was employed for genomic sequences expected to contain the full transcript and was helpful due to the tails being less well conserved and hence this choice aided in establishing the regions of lower identity or similarity as still being homologous.

The specificity of GeneWise has been reported as 87-96%, dependent on phylogenetic distance. Patterns in the false positives were low complexity regions and alternative splicing.\textsuperscript{(79)} Alternative splicing is possible in the \textit{H. contortus} P-gp and could have generated spurious matches but the model was robust enough. In this study the sequences had already been through a selection process and so specificity was less of a concern.

The sensitivity of GeneWise was relatively low in previous studies with only 60% of known transcripts being detected if the input is a short protein or a fragmented amino acid sequence. With the long and full-length P-gp sequences from \textit{C. elegans} this was unlikely to be a problem. In other cases the sensitivity was reported to rise to 98% when the similarity of the amino acid sequences was 85-95% through to within 20 bases of the termini.\textsuperscript{(79)} The incomplete \textit{H. contortus} genome and hence fragmented genes was a potentially limiting factor in the sensitivity.

All sequences from the S09 database that had previously been aligned with \textit{pgp} were re-evaluated using this GeneWise algorithm. The genomic sequences were entered as the target and the homologous amino acid sequence entered as the query into the \textit{GeneWise} component of the WISE2 package hosted at Mobyle (mo byle.pasteur.fr/cgi-bin/portal.py?form=wise2). The comparison matrix used was BLOSUM62 and the gene parameter file selected was \textit{worm}. The type of match attempted was variously \textit{global} or \textit{local} dependent on which predicted the longest gene structure with the least artefacts. Likewise the figures used as gap penalties.
and gap extension penalties were altered from the defaults of 12 and 2 to provide better results as necessary. The substitution and insertion/deletion error rates were rarely altered from the defaults of $10^{-5}$. All parameters have been provided at the start of every result, included in the Appendix. For each putative *pgp* sequence the predicted alignment, cDNA and amino acid sequences were generated; both the positions and sequences of the introns and the exons were provided. The known lengths of unknown sequence, denoted as strings of $n$, in the *supercontigs* did cause artefacts in the GeneWise outputs as inappropriate cut-offs in exons or were translated into unknown residues, denoted as $X$. Longer gaps in some *supercontigs* prevented resolution until the sequence was analysed in two sections and then manually edited.
3 RESULTS

3.1 Alignments along Entire Gene

3.1.1 Alignments from Assembled Contigs and Expressed Sequence Tags Databases

The first stage of the bioinformatic approach taken in this research is of querying the incomplete *H. contortus* genome, via the multiple databases becoming available over the course of the research, with known Cel-PGP. This produces high-scoring segment pairs (HSPS) with an example in Figure 1.
Figure 1: Alignments of *H. contortus* Sequences from CON Database with *Cel*-PGP-4.

Residue position of *Cel*-PGP-4 along x axis. Alignment of HSPs from both the forward and complementary sequences from CON database. Separation between matches indicates the fragmented short genome coverage as exons and separating introns are included in the bars. Marked increase of alignments at the ABC domains both 5’ (around residue position 500) and 3’ (around residue position 1150) indicate the conservation of these domains and alignments from genomic sequence homologous to other members of the *pgp* family.

The CON database of assembled *H. contortus contigs* released in 2006 revealed multiple matches on searching with BLAST. Figure 1 illustrates the HSPS matching *Cel*-PGP-4, and is representative of the other 14 members of the P-gp family in *C. elegans*. The *contigs* had multiple HSPS therefore the number of HSPS in Figure 1 is greater than the number of *contigs*, this is also demonstrated in Figure 2.
Examining the distribution of the HSPS in Figure 1 clearly shows the cluster around amino acid position 350-600 and from 1000 to the carboxyl-terminus (C-terminus), these are the two ABC and are highly conserved enabling highly scored alignments. This increases the sensitivity of this method to detect homologous sequences in these regions. This figure is similar to the graphic output from querying the other databases, although the number of HSPS varies. The later databases show increasing numbers of HSPS and greater coverage of the pgp family.

3.1.2 Alignments from Eight Databases

To visualise the extent of coverage of the *H. contortus* genome by the various Sanger databases the aligned *H. contortus* regions were depicted on a floating bar chart in Figure 2. The positions of these *H. contortus* sequences within each putative *H. contortus* PGP protein sequence were by analogy with the annotated locations of the homologous *C. elegans* sequences within the relevant *C. elegans* PGP protein. Figure 2 is an important chart as it illustrates the significant differences between the earliest databases in terms of the length and number of sequences returned, which is summarised in Table 5. The FOS database produces no pgp sequences. The BAC database gives only one relatively short sequence. The UNA source reveals matches along most of the length of pgp but with none overlapping and therefore provides no evidence of more than one pgp. The SEQ databases (SEQ1, SEQ2 and SEQ3) show increasing coverage with later releases as is expected from better and more thorough genome coverage. The CON database provides the greatest number of alignments at this stage, which were grouped in terms of which Cel-PGP they were most homologous to and these were then plotted in separate rows in the bar chart, one of these being the known mRNA and protein having been named pgp-A and PGP-A respectively. This early evidence points to ten Hco-pgp but later evidence in

Figure 8 is for nine pgp in *H. contortus*. However, the alignments these scores are calculated on are very short with just a few dozen amino acids from a couple of hundred nucleotides. The short white gaps in the bars indicate lack of alignment at the start and end of exons and are artefacts from the BLAST method rather than missing sequences. Following on from the CON database, the improved assembled contigs database (ASS; released in November 2007) containing longer
sequences is investigated in Figures 3-10 with alternative approaches to basic local alignment.

Figure 2: Alignments of H. contortus Sequences from Multiple Databases with Homologous Cel-PGP.

Residue positions of homologous Cel-PGP along x axis. Bars: BLAST alignments for sequence from H. contortus genomic databases to any Cel-PGP. No alignments from FOS database. BAC database provided 1 sequence homologous to Cel-PGP. UNA database provided fragmented but not overlapping sequences with no evidence for multiple genes. SEQ1, SEQ2 and SEQ3 databases provided fragmented short coverage but less fragmented longer sequences with each database release. SEQ databases provided no overlapping sequences and hence no evidence for more than 1 gene. CON database provided fragmented alignments with some overlap providing evidence for 10 genes. Marked fragmentation of genomic sequencing is evident in these long genes. Compare to greater alignments with the later databases C09 in Figure 9 and S09 in Figure 11.

Evidence for the number of pgp genes from the various databases is summarised below in Table 5. The FOS database provides no evidence of pgp, in the next five releases there is coverage but no overlapping sequences and therefore no more than one gene is demonstrated. The CON database contains 10 overlapping pgp sequences.
### Table 5: Evidence for the Number of *pgp* Genes in Different Databases.

<table>
<thead>
<tr>
<th>Database</th>
<th>Number of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOS</td>
<td>0</td>
</tr>
<tr>
<td>BAC</td>
<td>1</td>
</tr>
<tr>
<td>UNA</td>
<td>1</td>
</tr>
<tr>
<td>SEQ1</td>
<td>1</td>
</tr>
<tr>
<td>SEQ2</td>
<td>1</td>
</tr>
<tr>
<td>SEQ3</td>
<td>1</td>
</tr>
<tr>
<td>CON</td>
<td>10</td>
</tr>
<tr>
<td>Maximum</td>
<td>10</td>
</tr>
</tbody>
</table>

### 3.2 Alignments to Specific Domains

CD-Search reveals which conserved domains are present in the query and enables consistent comparison and positioning of different sequences from both *H. contortus* and *C. elegans*. Figure 3 illustrates the matches from the CD-Search on one *H. contortus* contig. Characteristic for the ABC within P-gps is the ATP binding site and ABC transporter signature. The entire ABC is recognised by the conserved domain (cd) *cd03249*. The sequence also registers with an ABC membrane that is registered as *pfam00664* and represents the six transmembrane helices of ABC transporters (the TMD). Under the multi-domain hits is the pattern registered in the clusters of orthologous groups (COG) database as *COG1132* for both the TMD and ABC. The other interesting hit is also a multi-domain model named *PTZ00265* and derives from the multidrug resistance (MDR) protein in *Plasmodium*. There are a great number of motif and domain matches not included in Figure 3.
Figure 3: Motif and Domain Alignments from CD-Search to a H. contortus contig from ASS Database.

Residue position of H. contortus contig 0009203 along x axis. Motif matches include the ATP binding site and ATP transporter signature. Domain matches include multiple ABC domain models including cd03249. Superfamily matches include the ABC membrane pfam00664. Multi-domain matches include the TMD and ABC model COG1132 and the MDR pattern PTZ00265. Matches indicate strong signal for ABC in contig 0009203.

3.2.1 Alignments to ATP Binding Cassettes

The ABC model (cd03249) is used to match contigs to this domain in Figures 4-8. The ABC (cd03249), TMD (pfam00664) and multi-domain models (COG1132 and PTZ00625) are compared in Figure 9 and then COG1132 is used to match contigs to both the TMD and ABC in Figure 10.

<table>
<thead>
<tr>
<th>Domain Position</th>
<th>Consensus Identity</th>
<th>Query Identity</th>
<th>Query Position</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>131</td>
<td>T</td>
<td>T</td>
<td>5</td>
<td>+7</td>
</tr>
<tr>
<td>132</td>
<td>L</td>
<td>K</td>
<td>6</td>
<td>-1</td>
</tr>
<tr>
<td>133</td>
<td>V</td>
<td>I</td>
<td>7</td>
<td>+2</td>
</tr>
<tr>
<td>134</td>
<td>G</td>
<td>G</td>
<td>8</td>
<td>+7</td>
</tr>
<tr>
<td>135</td>
<td>E</td>
<td>E</td>
<td>9</td>
<td>+4</td>
</tr>
<tr>
<td>136</td>
<td>R</td>
<td>G</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>137</td>
<td>G</td>
<td>G</td>
<td>11</td>
<td>+6</td>
</tr>
<tr>
<td>138</td>
<td>S</td>
<td>V</td>
<td>12</td>
<td>-1</td>
</tr>
<tr>
<td>139</td>
<td>Q</td>
<td>Q</td>
<td>13</td>
<td>+6</td>
</tr>
<tr>
<td>140</td>
<td>L</td>
<td>L</td>
<td>14</td>
<td>+6</td>
</tr>
</tbody>
</table>

Figure 4: Alignment of a H. contortus contig from ASS database with PSSM of ABC model.

Residue position of ABC domain model cd03249 in first column. Consensus residue identities of cd03249 in second column. Residue identities of contig 0009203 in third column. Residue position of H. contortus contig 0009203 in fourth column. Score from PSS Matrix of cd03249 for the residue in contig 0009203 in fifth column. Positive scoring for this portion of the domain model just prior to the ABC transporter motif (shown in Figure 5) indicates strong signal for ABC in contig 0009203.

The H. contortus sequence is aligned by RPS-BLAST to the PSSM in Figure 4 and the high scores for this alignment in the fifth column is evidence for the contig covering this part of the ABC.
<table>
<thead>
<tr>
<th>Domain Position</th>
<th>Consensus Identity</th>
<th>Query Identity</th>
<th>Query Position</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>140</td>
<td>L</td>
<td>L</td>
<td>14</td>
<td>+6</td>
</tr>
<tr>
<td>141</td>
<td>S</td>
<td>S</td>
<td>15</td>
<td>+6</td>
</tr>
<tr>
<td>142</td>
<td>G</td>
<td>G</td>
<td>16</td>
<td>+7</td>
</tr>
<tr>
<td>143</td>
<td>G</td>
<td>G</td>
<td>17</td>
<td>+7</td>
</tr>
<tr>
<td>144</td>
<td>Q</td>
<td>Q</td>
<td>18</td>
<td>+8</td>
</tr>
<tr>
<td>145</td>
<td>K</td>
<td>K</td>
<td>19</td>
<td>+7</td>
</tr>
<tr>
<td>146</td>
<td>Q</td>
<td>Q</td>
<td>20</td>
<td>+8</td>
</tr>
<tr>
<td>147</td>
<td>R</td>
<td>R</td>
<td>21</td>
<td>+8</td>
</tr>
<tr>
<td>148</td>
<td>I</td>
<td>V</td>
<td>22</td>
<td>+2</td>
</tr>
<tr>
<td>149</td>
<td>A</td>
<td>A</td>
<td>23</td>
<td>+6</td>
</tr>
</tbody>
</table>

**Figure 5: Alignment of a *H. contortus* contig from ASS database with ABC Transporter Motif in ABC Domain Model.**

Residue position of ABC transporter motif in domain model *cd03249* in first column. Consensus residue identities of *cd03249* in second column. Residue identities of *contig* 0009203 in third column. Residue position of *H. contortus* contig 0009203 in fourth column. Score from PSS Matrix of *cd03249* for the residue in *contig* 0009203 in fifth column. Exact match to consensus sequence at 9 of the 10 residues indicates strong signal for ABC transporter in *contig* 0009203.

Whilst a section of the alignment is shown in Figure 4, a more specific view is shown in Figure 5 with only the residues comprising the signature motif of the ABC transporter. The identity of this *contig* as an ABC transporter is demonstrated by the very high conservation to the signature motif in this figure.
Figure 6: Alignments of *H. contortus* contigs from ASS database with ABC Domain Model.

Residue position of ABC domain model *cd03249* along x axis. Full-length of domain model is 238 residues. Fragmented coverage is evident with only 6 contigs covering entire ABC. Alignments for 6 contigs start or end at the same residue position of approximately 120 in the domain model, likely to be a splice site conserved across genes.

The alignments generated between the ABC and the *H. contortus* contigs are plotted in Figure 6. Of the 62 contigs that were most similar to *pgp* sequences in a reciprocal BLAST, 33 aligned with the ABC. Only six contigs were aligned with the entire length of the domain and the remaining 27 contigs matched part of the domain. The number of overlapping sequences is clearest at the 5’ and 3’ ends where 18 contigs overlap, providing evidence for no more than nine *pgp* in *H. contortus*. 

44
Figure 7: Neighbour-Joining Tree of a *H. contortus* contig from ASS Database with the Aligned Portion of ABC Domain from *Cel-PGP*.

*ASS_0009203_r* is the complementary sequence (r: reverse) of *contig* 0009203 from the ASS database. N: N-terminus ABC domain. C: C-terminus ABC domain. *Cel-pgp-15ps_5*: 5' ABC domain of pseudo gene *Cel-pgp-15ps*. Numbers: number of trees of 1 000 replicates that divide the sequence into that same division. A ratio of 431 trees of 1 000 for homology of *contig* 0009203 with the N-terminus of the subgroup *Cel-PGP-12*, 13, 14 and 5' *Cel-pgp-15ps*. This is an unrooted tree.

The 33 *H. contortus* sequences from the ASS database with sequence similarity to the ABC domain were aligned with the ABC and this domain was used as a common positioning system. Both 5' and 3' domains of all *Cel-PGP* and one individual *H. contortus* translated sequence were aligned and a phylogenetic analysis carried out with the result of a neighbour-joining tree as in Figure 7. The numbers at each branch represents the number of times out of 1 000 bootstrapped repetitions this separation of the tree was generated. The *H. contortus* sequence in Figure 7 clusters with a particular *C. elegans* set of domains (the N-terminal ABC of the subgroup *Cel-PGP-12* to 15) in 431 of the 1 000 bootstrapped repetitions, which is very low and does not confirm this homology. However, this contig only aligned with 110 amino acids of the ABC
and hence any further analysis is hampered. When the contig that putatively covers the remainder of the ABC is included in the alignment the result rises to 93%. The phylogenetics assists in producing

Figure 8.

Figure 8: Alignment of *H. contortus* contigs from ASS Database Aligned with 5' portion of ABC Domain Model.

Residue position of ABC domain model *cd03249* along the top. Hyphens: gaps between *contigs*. Letters separated by hyphens: *contigs* that have been edited by the author to match the ABC with another *contig*. Other *contigs* have been matched by the author with no exonic gap and hence no hyphens. Dual number in identity key below: *contigs* combined by author. Reverse in identity key below: complementary sequence of *contig*. Clustal-X colour coding highlights residue positions of physico-chemical conservation. Amino acid conservation in ABC domain demonstrated by colour conservation. 1: 0012003 N-terminus. 2: 0012003 C-terminus. 3: 0103012 reverse. 4: 0006263. 5: 0089801. 6: 0019024. 7: 0019023 & 0103014. 8: 0007519 reverse & 0044039. 9: 0051131 reverse & 0102484. 10: 0089802 & 0043828. 11: 0061208 reverse & 0009202 reverse. 12: 0043738 reverse & 0092554. 13: 0103015 reverse & 0009203 reverse. 14: 0084658 & 0020529. 15: 0006158 reverse & 0050776 reverse. 16: 0020528 & 0043738 reverse & 0092554. 17: 0019023 & 0103014. 18: 0007519 reverse & 0044039.

Figure 8 is the end result of the *contigs* from the ASS database being aligned with the ABC. Examining how the *contigs* could possibly fit together and then taking account of the phylogenetic support for these combinations leads to these putative 18 sequences of the ABC in *H. contortus*. There are nine pairs and so is evidence for the presence of nine *pgp* in *H. contortus* as opposed to the earlier evidence for ten *pgp* as in Figure 2. The short gaps are between the end of one *contig* and the start of another. Other *contigs* are joined end-to-end with no gaps due to the break occurring within an intron. One *contig* covers a pair of ABC, this is 0012003 and includes the entire sequence of the previously discovered *Hco-pgp-2.1.A* that is renamed by the author from *Hco-pgp-A*. Four *contigs* cover an entire ABC domain each, while 11 domains are covered by a pair of *contigs*, some with exonic
gaps and some without and finally one domain is represented by four contigs. The methodology represented by the previous figures resulted in as great a coverage of the ABC domains as possible given the extent of coverage of the genome in the ASS database.

3.2.1.1 Alignments to Both ATP Binding Cassettes and Transmembrane Domains

![Figure 9: Coverage of Hco-PGP-2.1.A to ABC and TMD Domain Models.](image)

Figure 9: Coverage of Hco-PGP-2.1.A to ABC and TMD Domain Models.

Residue position of full-length Hco-PGP-2.1.A along x axis. Numbers in graph: length of alignments or gaps for this H. contortus sequence to the domain model. cd03249: domain model of ABC. pfam00664: domain model of TMD. COG1132: domain model of single TMD and ABC. PTZ00625: domain model of duplicated TMD and ABC. Successful alignment of each domain model to Hco-PGP-2.1.A is shown by bars of colour.

To extend the alignment of contigs beyond the ABC other domain models are compared. The coverage of these different models through their PSSM is shown in Figure 9. The longer coverage in PTZ00625 initially appears more useful, however it is based on Plasmodium and is not sensitive enough to pick up both 5' and 3' domains but instead disregards one domain and provides a shorter alignment of the contigs than COG1132. This COG1132 has as high a sensitivity to H. contortus sequences as the single domain matrices of cd03249 and pfam00664 whilst improving the coverage of the pgp and so is used to generate the alignments in Figure 10.
Figure 10: Alignment of *H. contortus contigs* from ASS Database at the Intersection of TMD and ABC Domains.


Of the 62 contigs that aligned with *Cel-pgp* on a reciprocal BLAST, 33 covered the ABC and a total of 53 are aligned within the COG1132 domain model incorporating both the TMD and ABC. Those contigs aligned within the COG1132 domain at the intersection of the TMD and ABC are shown in Figure 10. The alignment clearly shows a greater number of alignments on the right (18) than on the left (12) and this number decreases further to the 5’ end of the TMD, beyond the region shown in Figure 10 where the lower conservation decreases the sensitivity of the method to detecting contigs that cover these regions as opposed to the ABC. In particular, sequences that only covered the TMD were not efficiently detected.

Table 6: Evidence for the Number of *ppg* Genes in ASS Database.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Number of Domains</th>
<th>Number of Bi-domain Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABC</td>
<td>18</td>
<td>9</td>
</tr>
<tr>
<td>TMD</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>Maximum</td>
<td>18</td>
<td>9</td>
</tr>
</tbody>
</table>

The evidence for the number of *ppg* genes is summarised in Table 6. The ABC domain is present 18 times in the database and hence is evidence for 9 genes,
whereas the TMD is only found in 12 sequences and does not support more than 6 genes.

Figure 11: Alignments of *H. contortus* contigs from C09 Database with Homologous Cel-PGP.

Residue position of homologous Cel-PGP along x axis. Putative ABC signal is visible from the number of alignments around residue position 550 and 1150. Majority of contigs are relatively short with larger gaps; only one contig putatively covers both ABC.

The contigs that were established as homologues to pgp by reciprocal BLAST are plotted in Figure 11 showing the regions of alignment. There is varying evidence for the number of pgp genes over the length of the putative genes from nil start signals to 12 close to the 3’ end as summarised in Table 7.

Table 7: Evidence for the Number of pgp Genes in C09 Database.

<table>
<thead>
<tr>
<th>Position</th>
<th>Number of Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>0</td>
</tr>
<tr>
<td>550</td>
<td>10</td>
</tr>
<tr>
<td>1100</td>
<td>9</td>
</tr>
<tr>
<td>1200</td>
<td>12</td>
</tr>
<tr>
<td>Maximum</td>
<td>12</td>
</tr>
</tbody>
</table>
The C09 contigs are generally longer than those in the ASS database but are still too short for pgp full-length alignment therefore the next stage was the use of supercontigs as in the figure below.

### 3.3 Alignments along Entire Gene

![Diagram](image)

**Figure 12: Dot Plot of H. contortus supercontig from S09 Database Aligned with Cel-PGP-4.**

Residue position of Cel-PGP-4 along x axis and base position of H. contortus supercontig 001880 up y axis. Dots and lines: regions of identity with putative exons separated by vertical jumps up the y axis according to putative intron length. Direction from bottom left to top right indicates agreement to the provided sequence rather than complementary. Pattern suggests alignment from residue 100 to the stop signal from base 15 000 to 24 000 approximately. Characteristic 4-cornered box of bi-domain signal due to 5' H. contortus putative ABC domain matching both 5' and 3' C. elegans ABC domain and the 3' H. contortus putative ABC domain matching both 5' and 3' C. elegans ABC domain.

The HSPS generated by BLAST between a H. contortus supercontig and the homologous Cel-PGP can be plotted as in Figure 12. The conservation of the 5' and 3' ABC is apparent in the characteristic mirrored HSPS, resulting in the appearance of four matches from two genuine matches because the N-terminal ABC of the C. elegans pgp is aligned with both the 5' and 3' ABC of the H. contortus and the same for the C-terminal ABC in C. elegans. There is a lack
of alignment to the amino-terminus (N-terminus) and in the hypervariable cytoplasmic loop between the pair of domains using these methods. This was repeated for the other 14 supercontigs that contained HSPS to Cel-PGP. The alignment most likely to be true is the longest diagonal line, in this case it has major interruptions and the long gaps remain even within highly conserved domains suggesting the sensitivity of the alignment software is not at fault. The annotated supercontig in Sequence 1 below shows the reason for these gaps.

```
3.3.1 Gene Structure Prediction Algorithms

The successful gene structure prediction algorithm in this study was the GeneWise algorithm. The output for the same supercontig as in

Sequence 1: Annotated *H. contortus supercontig* 0001880 from S09 Database.

Full nucleotide sequence of supercontig aligned to *C. elegans* protein. Yellow highlight: unknown sequence between contigs within supercontig. Red highlight: *C. elegans* amino acid sequence unmatched.

All supercontigs implicated in the preceding stages were investigated using the BLAST method but only one example is given above. The sections of unknown sequence and regions of hypervariability were handicaps to determining full-length nucleotide and amino acid sequence as shown by the regions of red and yellow highlights so an improvement was sought using gene structure prediction algorithms as shown below in Sequence 2.
Sequence 1 is shown in Sequence 2 below. The entire set of outputs for the 48 supercontigs from the S09 database with evidence of pgp exons are in the Appendix.

<table>
<thead>
<tr>
<th>Query protein:</th>
<th>CAEEL-PGP-4.1a</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp Matrix:</td>
<td>blosum62.bla</td>
</tr>
<tr>
<td>Gap open:</td>
<td>12</td>
</tr>
<tr>
<td>Gap extension:</td>
<td>2</td>
</tr>
<tr>
<td>Start/End:</td>
<td>global</td>
</tr>
<tr>
<td>Target Sequence:</td>
<td>H_S09_0001880__ S09_Supercontig 0001880</td>
</tr>
<tr>
<td>Strand:</td>
<td>both</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene wise</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Query protein:</td>
<td>CAEEL-PGP-4.1a</td>
</tr>
<tr>
<td>Comp Matrix:</td>
<td>blosum62.bla</td>
</tr>
<tr>
<td>Gap open:</td>
<td>12</td>
</tr>
<tr>
<td>Gap extension:</td>
<td>2</td>
</tr>
<tr>
<td>Start/End:</td>
<td>global</td>
</tr>
<tr>
<td>Target Sequence:</td>
<td>H_S09_0001880__ S09_Supercontig 0001880</td>
</tr>
<tr>
<td>Strand:</td>
<td>both</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene wise</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Query protein:</td>
<td>CAEEL-PGP-4.1a</td>
</tr>
<tr>
<td>Comp Matrix:</td>
<td>blosum62.bla</td>
</tr>
<tr>
<td>Gap open:</td>
<td>12</td>
</tr>
<tr>
<td>Gap extension:</td>
<td>2</td>
</tr>
<tr>
<td>Start/End:</td>
<td>global</td>
</tr>
<tr>
<td>Target Sequence:</td>
<td>H_S09_0001880__ S09_Supercontig 0001880</td>
</tr>
<tr>
<td>Strand:</td>
<td>both</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene wise</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Query protein:</td>
<td>CAEEL-PGP-4.1a</td>
</tr>
<tr>
<td>Comp Matrix:</td>
<td>blosum62.bla</td>
</tr>
<tr>
<td>Gap open:</td>
<td>12</td>
</tr>
<tr>
<td>Gap extension:</td>
<td>2</td>
</tr>
<tr>
<td>Start/End:</td>
<td>global</td>
</tr>
<tr>
<td>Target Sequence:</td>
<td>H_S09_0001880__ S09_Supercontig 0001880</td>
</tr>
<tr>
<td>Strand:</td>
<td>both</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene wise</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Query protein:</td>
<td>CAEEL-PGP-4.1a</td>
</tr>
<tr>
<td>Comp Matrix:</td>
<td>blosum62.bla</td>
</tr>
<tr>
<td>Gap open:</td>
<td>12</td>
</tr>
<tr>
<td>Gap extension:</td>
<td>2</td>
</tr>
<tr>
<td>Start/End:</td>
<td>global</td>
</tr>
<tr>
<td>Target Sequence:</td>
<td>H_S09_0001880__ S09_Supercontig 0001880</td>
</tr>
<tr>
<td>Strand:</td>
<td>both</td>
</tr>
</tbody>
</table>

...
ggctacgacacaatatggaagggctcttacgccggacaaacctcacgagacgaagcagccagctcgtggccgctagcggcagttgctatcgctcgagccctggttaggaatccacgaatacttctactggatgaggcaacgagtgctttggatgcagaaagcgaaagtaa)

//
query_protein: CAEEL-4_1a_4_1a_4
comp_matrix: blosum62.bla
gap_open: 1
gap_extension: 2
start/end local

target_sequence: H_S09_0001880___

//
genewise

query protein: CAEEL-4_1a_4_F42E11_1a_1-1280AA
comp_matrix: blosum62.bla

//
GeneWise parameters: previously established as homologous protein for the putative transcript from *H. contortus supercontig* 0001880. Global model parameter in GeneWise generated greatest alignment. Alignment starts an estimated 47 codons from start; residues form positions 271-514 and 1072-1167 not matched; alignment extends to stop signal. Alignment of *C. elegans* residues, translated codons and nucleotides provided then list of base positions of exons, list of introns in phase 1 or 2, split codons, translation of putative *H. contortus* sequence, transcript of putative *H. contortus* sequence. Additional alignment to *C. elegans* residue positions 1070-1163 provided, despite including a frame shift. Green highlight: alignment. Yellow highlight: unknown sequence between contigs within supercontig.

GeneWise picked up two additional exons (starting at nucleotide position 13981 in Sequence 2) 5' to the first exon (starting at nucleotide position 14795 in Sequence 1) detected by the BLAST method. However, the start of the gene was not detected by either method. GeneWise was also sensitive to the exon starting at position 15690, missed by BLAST. Where no identities existed for a stretch of 884 and then 480 nucleotides no method could be effective. GeneWise predicted an exon (starting at position 18384) with weak homology and two strings of 8 and 9 residues in the *C. elegans* sequence with no alignment to *H. contortus* whilst the BLAST method just indicates a lack of alignment. This is in the region between

Sequence 2: Annotated *H. contortus supercontig* 0001880 from S09 Database using the GeneWise Algorithm.
domains that does commonly demonstrate variable length between proteins and species.

The number of sequences from the S09 database that were aligned with E-values below 10, was between 57 (against Cel-PGP-11) and 73 (against Cel-PGP-2). The sequence on the forward BLAST that was lowest in the order of E-values that still provided evidence of pgp coverage on the back BLAST was the 47th sequence out of 67 sequences (against Cel-PGP-10). This provided evidence that an E-value of 10 as a cut-off point was still sufficiently sensitive. The total number of sequences that were aligned to pgp genes with an E-value below 10 during the forward BLAST was 95. The number with better alignment to pgp rather than other gene families (over part or all of the sequence) on the back BLAST was 48. The remaining 47 sequences were considered by this evidence to be false positives from the forward BLAST.

Manually editing the alignments generated by BLAST from the S09 database available during this final phase and then plotting the alignments for the supercontigs results in Figure 13. This view demonstrates the lack of detected alignments to the N-terminus. Also Figure 13 illustrates not just the maximal length but does highlight in yellow the substantial gaps within the stretch. Not only are there gaps within the supercontig coverage of the pgp but even with this latest database there are only two pgp fully covered by a single supercontig each.
Figure 13: Alignments of *H. contortus* supercontigs from S09 Database with Homologous Cel-PGP.

Scale on x axis in amino acids of the homologous Cel-PGP; hence the STOP codon at different locations. 2.1.A is *Hco*-PGP-2.1.A. Solid black bars: alignment. Vertical bars: alternative splices. Solid grey bars: unknown sequence between contigs within supercontigs. Red bars with cap: stop codons. Order is first by homologous single Cel-PGP or subgroup and second by 5' position of alignment. Each row is a single supercontig; supercontig number is in Table 8 below.
The detail of the alignments in Figure 13 are summarised in Table 8 with just the earliest start and latest end of the multiple and alternative alignments to the amino acid positions in the homologous Cel-PGP. The rows are in order and so indicate the number of the *H. contortus* supercontig for the applicable bar in Figure 13.
Table 8: Summarised Alignments of *H. contortus supercontigs* from S09 Database with Homologous Cel−PGP.

<table>
<thead>
<tr>
<th>Homologous Cel−PGP</th>
<th><em>H. contortus supercontig</em></th>
<th>Earliest Amino Acid Aligned</th>
<th>Last Amino Acid Aligned</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0050550</td>
<td>1</td>
<td>109</td>
</tr>
<tr>
<td>1</td>
<td>0045305</td>
<td>157</td>
<td>226</td>
</tr>
<tr>
<td>1</td>
<td>0034154</td>
<td>228</td>
<td>301</td>
</tr>
<tr>
<td>1</td>
<td>0061621</td>
<td>301</td>
<td>515</td>
</tr>
<tr>
<td>1</td>
<td>0006074</td>
<td>543</td>
<td>1321</td>
</tr>
<tr>
<td>1</td>
<td>0032856</td>
<td>627</td>
<td>665</td>
</tr>
<tr>
<td>1</td>
<td>0011236</td>
<td>730</td>
<td>984</td>
</tr>
<tr>
<td>1</td>
<td>0044638</td>
<td>869</td>
<td>984</td>
</tr>
<tr>
<td>1</td>
<td>0046186</td>
<td>985</td>
<td>1178</td>
</tr>
<tr>
<td>1</td>
<td>0047549</td>
<td>1162</td>
<td>1321</td>
</tr>
<tr>
<td>1</td>
<td>0024971</td>
<td>1266</td>
<td>1321</td>
</tr>
<tr>
<td>2</td>
<td>0049649</td>
<td>394</td>
<td>525</td>
</tr>
<tr>
<td>2.1.A</td>
<td>0007774</td>
<td>40</td>
<td>875</td>
</tr>
<tr>
<td>2.1.A</td>
<td>0038725</td>
<td>1025</td>
<td>1158</td>
</tr>
<tr>
<td>2.1.A</td>
<td>0046285</td>
<td>1177</td>
<td>1275</td>
</tr>
<tr>
<td>2.1.A</td>
<td>0005974</td>
<td>1208</td>
<td>1275</td>
</tr>
<tr>
<td>3</td>
<td>0001706</td>
<td>117</td>
<td>822</td>
</tr>
<tr>
<td>3</td>
<td>0059287</td>
<td>196</td>
<td>260</td>
</tr>
<tr>
<td>3</td>
<td>0023983</td>
<td>823</td>
<td>890</td>
</tr>
<tr>
<td>3/4</td>
<td>0000863</td>
<td>943</td>
<td>1242</td>
</tr>
<tr>
<td>3/4</td>
<td>001880</td>
<td>47</td>
<td>1280</td>
</tr>
<tr>
<td>3/4</td>
<td>001880</td>
<td>1070</td>
<td>1099</td>
</tr>
<tr>
<td>4</td>
<td>0057179</td>
<td>1135</td>
<td>1280</td>
</tr>
<tr>
<td>9</td>
<td>0006925</td>
<td>205</td>
<td>1294</td>
</tr>
<tr>
<td>9</td>
<td>0006925</td>
<td>1</td>
<td>1294</td>
</tr>
<tr>
<td>9</td>
<td>006925</td>
<td>557</td>
<td>821</td>
</tr>
<tr>
<td>9</td>
<td>0038098</td>
<td>35</td>
<td>1269</td>
</tr>
<tr>
<td>9</td>
<td>0046372</td>
<td>94</td>
<td>225</td>
</tr>
<tr>
<td>9</td>
<td>0046372</td>
<td>676</td>
<td>731</td>
</tr>
<tr>
<td>9</td>
<td>0016882</td>
<td>1090</td>
<td>1294</td>
</tr>
<tr>
<td>10</td>
<td>0035472</td>
<td>120</td>
<td>179</td>
</tr>
<tr>
<td>10</td>
<td>0025718</td>
<td>219</td>
<td>377</td>
</tr>
<tr>
<td>10</td>
<td>0064200</td>
<td>374</td>
<td>507</td>
</tr>
<tr>
<td>10</td>
<td>0038139</td>
<td>508</td>
<td>578</td>
</tr>
<tr>
<td>10</td>
<td>0037674</td>
<td>793</td>
<td>844</td>
</tr>
<tr>
<td>10</td>
<td>0037517</td>
<td>839</td>
<td>985</td>
</tr>
<tr>
<td>11</td>
<td>0024002</td>
<td>64</td>
<td>380</td>
</tr>
<tr>
<td>11</td>
<td>0024351</td>
<td>301</td>
<td>871</td>
</tr>
<tr>
<td>11</td>
<td>0035404</td>
<td>1047</td>
<td>1234</td>
</tr>
<tr>
<td>14</td>
<td>0069412</td>
<td>127</td>
<td>212</td>
</tr>
<tr>
<td>14</td>
<td>0004549</td>
<td>214</td>
<td>270</td>
</tr>
<tr>
<td>13</td>
<td>0059902</td>
<td>340</td>
<td>396</td>
</tr>
<tr>
<td>12</td>
<td>0021658</td>
<td>501</td>
<td>595</td>
</tr>
<tr>
<td>12</td>
<td>0050285</td>
<td>593</td>
<td>660</td>
</tr>
<tr>
<td>14</td>
<td>0015622</td>
<td>828</td>
<td>1079</td>
</tr>
<tr>
<td>14</td>
<td>0021166</td>
<td>1030</td>
<td>1217</td>
</tr>
<tr>
<td>13</td>
<td>0010288</td>
<td>1224</td>
<td>1324</td>
</tr>
<tr>
<td>14/15ps</td>
<td>0024244</td>
<td>702</td>
<td>1261</td>
</tr>
<tr>
<td>15ps</td>
<td>0005706</td>
<td>144</td>
<td>1020</td>
</tr>
<tr>
<td>1/5/6</td>
<td>0013014</td>
<td>1003</td>
<td>1252</td>
</tr>
<tr>
<td>4/9</td>
<td>0055664</td>
<td>733</td>
<td>788</td>
</tr>
<tr>
<td>5/12/14</td>
<td>0055790</td>
<td>608</td>
<td>646</td>
</tr>
</tbody>
</table>
Eleven stop codons were detected within sequences that showed homology to Cel-PGP and are indicated in Figure 13 as final capped red bars. However, several of these sequences were relatively short and multiple supercontigs were homologous to the same Cel-PGP (three with Cel-PGP-1 and three with Cel-PGP-9) without corresponding numbers of TMD and ABC. In multiple regions of multiple Cel-PGP there was more than one H. contortus supercontig aligned. The author concluded the evidence in Figure 13 supported a total of 11 H. contortus pgp genes as summarised in Table 9.

**Table 9: Number of pgp Genes in C. elegans and H. contortus.**

<table>
<thead>
<tr>
<th>Cel-PGP or Subgroup</th>
<th>Number of Genes in C. elegans</th>
<th>Number of Homologues in H. contortus</th>
<th>Difference H. contortus vs. C. elegans</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3/4</td>
<td>2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>5/6/7/8</td>
<td>4</td>
<td>0</td>
<td>-4</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>3</td>
<td>+2</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>12/13/14/15ps</td>
<td>3 &amp; ps</td>
<td>2</td>
<td>-1 / -2</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>14 &amp; ps</strong></td>
<td><strong>11</strong></td>
<td><strong>-3 / -4</strong></td>
</tr>
</tbody>
</table>

The pgp gene families are compared between C. elegans and H. contortus in the table above highlighting the lack of any evidence of homologues to the Cel-pgp-5, 6, 7, 8; the two additional pgp-9 genes and the fewer genes homologous to the Cel-pgp-12, 13, 14, 15ps subgroup with the bioinformatic evidence from this study.
4 DISCUSSION

This thesis is concerned with revealing the \textit{pgp} genes in \textit{H. contortus}. The aim of this research was to discover the incomplete and dispersed P-gp genetic sequences in the sequence databases and construct the putative genes. These could then be investigated in the future for their rôle in IVM resistance in \textit{H. contortus} and potentially utilised in molecular tests for IVM resistance.

In the introduction \textit{H. contortus} was described as a parasitic nematode of sheep and goats that feeds on blood in the stomach, which can lead to symptoms of anaemia, lethargy, weight loss and even death in acute infections. As it is one of the most prevalent and economically important parasites internationally, its treatment and control is paramount. The importance of this species was the reason for this study on it.

Treatment and control of helminths in general and \textit{H. contortus} in particular is dependent on anthelmintics as there are no practical alternatives, to the extent that anthelmintics constitute 40\% of all livestock health expenditure. Commercial introduction of IVM was in 1981 and has grown to become the most commonly used drug in livestock agriculture. The importance of this drug was the reason for this study being directed at its use.

The high dependence on anthelmintics in general and IVM in particular means resistance has serious outcomes. There is already widespread resistance to the older anthelmintics including narrow-spectrum and all three classic broad-spectrum classes of drug, which has increased the reliance on IVM. The emergence of IVM resistance in \textit{H. contortus} is therefore a serious threat to the control of the pathogen to the extent of threatening the entire sheep industry in Australia. The significance of IVM resistance in \textit{H. contortus} was the motivation behind this study into this particular combination.

Presently IVM resistance can be identified and measured on the basis of changes in phenotype. However, these phenotypic tests are not sensitive or reliable enough to detect early selection of resistance but are restricted to being indicators of clinical resistance. The genetic basis of IVM resistance in \textit{H. contortus}
is not currently known\(^{(6; 71)}\) which precludes any monitoring of resistant genotypes prior to phenotypic changes. The desire for genotypic tests for the benefit of academic research of resistance development and agricultural application of resistance control is the background for this thesis.

As the *H. contortus* genome has not yet been completed, and therefore not annotated, genome-wide approaches are currently not available for determining genes or loci under IVM selection that might serve as signals for genotype changes and tests. Therefore the approach taken to date is the analysis of individual genes or gene families implicated in IVM resistance. One candidate for the genetic route of IVM resistance is the P-gp family. There has been overwhelming evidence of their rôle in anthelmintic activity\(^{(60; 61)}\) and P-gps have been associated with ML resistance from population genetic studies;\(^{(31; 62)}\) although this was not consistent.\(^{(63)}\) In another parasitic nematode, selection on an ABC transporter (the superfamily that includes P-gps) was verified,\(^{(64; 65; 66; 67)}\) and in both *H. contortus* and the other nematode IVM treatment led to decreased polymorphism of P-gps, a classic indication of selection.\(^{(31; 68; 69)}\) This background suggested the P-gp gene family should be investigated in this current study.

The P-gp family is well studied in many pathogens due to their involvement in multiple drug resistance to the extent that some of these proteins are named Multiple Drug Resistant Protein. However, in *H. contortus* the phylogenetics and ontology of the family is not yet established and even the definitive number of genes is unknown. As already stated, the *H. contortus* genome has yet to be completed and there is therefore no annotation of genes available. Only one P-gp mRNA transcript has been fully sequenced\(^{(70)}\) and the other members have only partial sequences submitted, which are a fraction of their full length.\(^{(63; 68)}\) The gene sequences remain fragmented and dispersed in the sequence databases. The P-gps are very long, contain multiple repeats and the family consists of multiple members that are similar; hence they are not easily amenable to most bioinformatic approaches.

The *pgp* genes have a repeated pair of domains with a 5' TMD and ABC and a 3' TMD and ABD. These domains are therefore physically close on the chromosomes. However, duplicated genes are also physically close in *C. elegans*, for example: the two members in the sub-group of *Cel-pgp-3* and 4;
the four members of the sub-group of Cel-pgp-5, 6, 7 and 8 or the four members of the sub-group of Cel-pgp-12, 13, 14 and the pseudogene 15ps. Therefore a pair of domains on a single H. contortus genomic fragment could be the 5' and 3' domains of a single gene but may well be the 3' domain of one gene and the 5' domain of a second separate gene. As the region before the 5' domain as well as between the 5' and 3' domains is hypervariable, discerning whether a sequence is 5' or 3' is limited. Because the initial Met and exon can be a long intron distant from the subsequent exons, the signal from the start of a 5' domain is not necessarily strong. The stop signal is an absolute signal for a domain being 3' unless a potential splice site is present and the stop signal is intronic. A number of sequences did not continue to the 3' end of the domain or unknown nucleotides between contigs in the supercontig blocked the signal. In the case of sequences only containing one domain with no 3' coverage of the potential stop signal the domain could be 5' or 3' in a gene. The homologous Cel-PGP for sequences with bi-domain coverage and a clear 5' start or 3' stop signal was determined by full-length comparison to both domains of all members of the Cel-PGP. In this case the only question was which individual or sub-group was the H. contortus sequence homologous to. Those sequences with bi-domain coverage but without a clear 5' start or 3' stop signal were analysed by comparison of each domain separately to Cel-PGP. In those cases the domains were independently homologous to 5' one and 3' for the other. In no case was a 3' domain contiguous with a 5' domain downstream in the sequence. In conclusion, no bi-domain sequences showed evidence of covering adjacent genes but did indicate that a single gene was present.

In the scenario of sequences only containing one domain with no 3' coverage of the potential stop signal, the domain could be 5' or 3' in a gene and it could be one of several gene homologues. The 5' and 3' domains of Cel-PGP were provided separately for homology in order to differentiate between both 5' and 3' as well as gene homology.

The above points do not take into account confounding signals in the sequence. The level of sequence difference between 5' and 3' domains in individual genes is similar to that between 5' domains in different genes in the same sub-group and between 3' domains in different genes in the same sub-group in C. elegans and is considered to be the same situation for H. contortus. In this case only weak
homology to one of the domains in one of the genes can be expected and was indeed found.

In addition to clear sequence providing weakened evidence, polymorphism in the genome of nematodes, including exons, is high and in *H. contortus* is particularly high. The outcome is that a sequence with no clear signal to a particular homologue can be swayed in one direction or another by polymorphisms. Therefore the stated homologous gene and the domain being identified as 5’ or 3’ is more tentative due to polymorphism.

Despite these difficulties it is considered worthwhile researching the genetic sequences of the *pgp* family so that the sequences and expression can be compared between sensitive and resistant populations and thus to assess *pgp* involvement in IVM resistance. The purpose of the bioinformatics in this study is to discover novel genetic sequences in the incomplete *H. contortus* genome that map both from and to known P-gp sequences and can be constructed into putative *H. contortus pgp* genes. The methodology has involved a combination of brute force sequence alignment followed by advanced bioinformatic algorithms.

The first stage of the bioinformatic approach taken in this research is of querying the incomplete *H. contortus* genome, via the multiple databases becoming available over the course of the research, with known Cel-PGP. Examining both the distribution of the HSPS and the histogram labelled coverage in Figure 1 clearly shows the cluster around amino acid position 350–600 and from 1000 to the C-terminus, these are the two ABC and are highly conserved enabling highly scored alignments. This increases the sensitivity of this method to detect homologous sequences in these regions. In the same way that the level of nucleotide variability between genes and individuals confounds determination of genomic sequence it also restricts the usefulness of local alignments as in BLAST.

Despite the presence of polymorphisms the P-gps, in common with all proteins, contained functional residues that were highly conserved. In addition, whole regions have particular functions and the range of variation at each residue is limited if the function of that domain is to be retained and for there not to be a fitness and survival penalty. This was exploited in the alignment of genomic sequences to known domains. This approach was the most apt for the P-gp
family as synonymous SNPs were disregarded and non-synonymous SNPs were assessed according to the matrix of scores depending on the specific position of the variant amino acid. This analysis was considered to come closer to a test of function and so more reliable as an indicator of gene identity than the direct study of sequence. However, full resolution of the TMD was restricted due to the lower conservation in comparison to the ABC.

The sections of unknown sequence in the supercontigs are a handicap to determining the full-length nucleotide and amino acid sequence. Not only are there gaps within the supercontig coverage of the pgp family but even with the latest database there are only two pgp fully covered by a single supercontig each. Several of the aligned sequences were relatively short thus reducing the reliability of these alignments. The summary of the pgp gene family coverage in Figure 13 demonstrates the lack of detected alignments to the N-terminus. Multiple supercontigs were homologous to the same Cel-PGP (three with Cel-PGP-1 and three with Cel-PGP-9) without corresponding numbers of TMD and ABC, in particular towards the N-terminus. In multiple regions of multiple Cel-PGP there was more than one H. contortus supercontig aligned. Whilst alternative splicing has been evident within supercontigs, determining whether overlapping supercontigs are alternative splices has not been conclusive.

There are eight clades in C. elegans (Cel-pgp-1; 2; 3/4; 5/6/7/8; 9; 10; 11; 12/13/14/15) of which seven have representatives in H. contortus; there are no homologues to the entire clade of four genes Cel-pgp-5 to 8 inclusive. All the duplications in C. elegans are on the X chromosome but none of these groups has more than one homologue in H. contortus. Conversely autosomal Cel-pgp-2 and Cel-pgp-9 with evidence of multiple homologues in H. contortus are not duplicated in C. elegans.

Despite H. contortus diverging from C. elegans approximately 400 million years ago\(^{81}\) the striking aspect of the phylogenetics of the pgp members is the consistency with which H. contortus pgp cluster alongside C. elegans rather than the two species providing two distinct branches from the common root. One example of a neighbour-joining tree is shown in Figure 7 and other phylogenetic approaches demonstrated the same pattern. This is characterised as a deeply-rooted tree and means all of the intervening evolution has maintained the
homology within and comparative branching between each sub-group. Genes from the same sub-groups but different species have more similarity than genes from different sub-groups in the same species. Analysing just one *H. contortus* sequence removed any confounding potential of another *H. contortus* sequence changing the branching statistics. As the purpose of this comparison was to determine the identity of the *H. contortus* sequence in terms of which member or sub-group of the P-gp family and whether 5' or 3' domain, the tree included both species of *H. contortus* and *C. elegans* but was not rooted as the evolutionary distance was not required. Furthermore the sub-groups of the P-gp family are very deep rooted so that sub-group conservation is stronger than species differentiation. The inclusion of the entire bi-domain P-gp gene family from an outgroup species was considered to be an unnecessary complication.

A more positive aspect to this pattern of evolution is that it appears the *pgp* do not mutate quickly or substantially. With the assumption that the selective pressure on farms does not significantly alter this past record a simple diagnostic test for a resistant genotype could be designed that is not rapidly made void by mutations and indels. A quantitative measure of IVM resistance due to P-gp mechanism could thus be calculated. A different approach is the analysis of this family for *pgp*-specific targets as the deeply-rooted evolutionary trees suggest that these would remain effective over time.

The early evidence points to ten *pgp* genes but later evidence in Figure 8 is for nine *pgp* in *H. contortus* and the results generated from the GeneWise algorithm indicates the number of stop codons is eleven as shown in Figure 13. The *pgp* family has a relatively large membership of 11 (*H. contortus*) or 15 (*C. elegans*) that presumably reflects the common requirement for those xenobiotic functions provided. Conversely the *Cel-pgp*-5–8 clade would be conjectured to provide functions not required by the parasite or non-homologous proteins cover those ontologies. The substantial number of P-gps in the parasitic nematode provokes concern regarding the prevention or control of IVM resistance via P-gp mechanisms as multiple proteins may be capable of contributing to resistance under selective pressure. This scenario is particularly likely when *H. contortus* is exposed to multiple drugs and classes of anthelmintic.
The *in silico* approach taken in this thesis needs to be a rough sketch that an *in vitro* approach can take and provide the full picture. The sequences provided in full in the Appendix could be used for the design of PCR primers. The successful amplifications would provide proof of the primers existing on contiguous genetic lengths and sequencing would confirm or expand upon the predicted sequences in terms of the primary question of whether the **contigs** and **supercontigs** in the genomic databases are indeed contiguous *in vitro*. In addition evidence for or against the exons and introns predicted by the GeneWise algorithm could be provided. When mRNA would be used the exonic and by comparison the intronic fragments would be shown. If using gDNA then longer amplicons including the introns would confirm simply that the sequences constructed *in silico* at Sanger were correct. It may be that the portion of sequence the primers are designed upon are present, however, the intervening sequence is partially or wholly different *in vitro*. Whilst the gDNA cannot show exon-intron boundaries, splicing or splicing alternatives the amplicons based on mRNA templates would indeed provide evidence of these essential and important processes.

It is particularly important to become aware of alternative splices early in genotype studies as quantitative studies such as quantitative real-time PCR (qPCR) may either be based on only one of the alternative mature mRNA products and be misrepresenting (probably under-representing) the quantity and rate of transcription of the gene at rest or in response to stimuli. The other aspect is that a difference in quantity and rate of production of the alternative splices would not be detected if there are not separately measured amplicons for each alternative. This would be the case in both scenarios of an amplicon bridging over alternative splice sites and therefore detecting all alternatives or a primer site being located in only one of the alternative splices and thus the other splices, if and when produced, would not be amplified or detected. There could be up-regulation of the gene, only one of the splices or a differential up-regulation with one of the splices having a greater change than the other. A more dramatic and immediately interesting aspect of alternative splice regulation is if up-regulation of one splice and down-regulation of the other splice was seen. Where an alternative splice is possible the detection of each and every splice should be measured simultaneously, within each experiment for that gene.
The alternative splice potential is important in sequence differences between sensitive and resistant populations to ensure no mutations or indels are missed by not sequencing all alternative splices. On detection of a difference of regulation of at least one splice of one gene, the region upstream of the gene should be sequenced for mutations or indels between the sensitive and resistant populations that may be correlated with the differential regulation. The sequence difference could, in addition to being correlated, be the cause via interaction with regulatory and transcription factors.

An example of expression causing resistance has been the insertion of a transposable element into the regulatory region of the cytochrome P450 gene Cyp6g1 causing up-regulation of the transcript in D. melanogaster and resistance to multiple insecticides that has spread globally.\(^{(57)}\) IVM selected strains were associated with P-gp overexpression in H. contortus.\(^{(70)}\) The linker domain in a P-gp between the ABC domains interacts with α and β-tubulin. These tubulin molecules are possibly arranged in the microtubules during the interaction. A suggested flow of information occurs with the P-gp & tubulin interaction dependent and differing according to microtubule integrity. As such the organelle status is signalled to the P-gp molecules. A further suggestion is that this signal up-regulates P-gp production.\(^{(82)}\) The explanation for such a cascade being selected is that lowered cell defence from microtubule compromise is counteracted by the defence mechanism of permeable glycoprotein pumps expelling toxins across the cell membrane.

Sex-dependent selection of β-tubulin genotypes by IVM treatment was discovered with females being selected but not males over 3 years in the human parasitic nematode O. volvulus.\(^{(20)}\) However, selection on male β-tubulin genotypes did occur over 6 years. \(^{(69; 83)}\) The explanation may be selection is initially of the females only but then selection occurs in their progeny and is detected over longer time scales with the year-long life cycle of O. volvulus. Sex-dependent mortality was detected also in O. volvulus with an estimated decrease in the female population of 27.5%, double that of an estimated 13.5% in male nematodes.\(^{(84)}\) The effect of IVM was therefore greater against female worms and could exert greater selection pressure on the females. Sex-dependent expression was estimated to be 24.3 fold higher in females of a gene putatively identified as a GluCl in the human parasitic nematode Brugia malayi.\(^{(85)}\) As GluCl are thought to
be the main target of IVM\(^{(86)}\) in certain configurations this could be significant for the IVM mode of action and efficacy between the different sexes in \(B.\ maltay\). The expression was not measured during IVM treatment. The action of IVM is speculated to be more efficacious in the presence of more molecular targets and IVM to have a more pronounced outcome in female \(B.\ maltay\).

If sex-dependent expression, selection and mortality were consistently greater in females across the species; a scenario in \(H.\ contortus\) could be envisaged of higher innate expression of IVM molecular targets in females, thus increasing selection for heritable resistance mechanisms in the face of greater mortality. Sequence and expression studies should therefore distinguish between male and female subjects and report results for both separately.

Whilst the \emph{in vitro} work outlined above could discover a genotype of \(H.\ contortus\) associated with an observed phenotype of IVM resistance, it may not be the ultimate cause. Not withstanding the use to pure science of determining causal factors, a correlation has a potential rôle as a molecular test of IVM resistance in a \(H.\ contortus\) population. Even if later research shows the test to be one of correlation with a different causal factor it can be utilised in the meantime.

A clear use of a genetic test would be in addressing the acknowledged deficit in our knowledge and understanding of the factors necessary or advantageous for both the initial development and subsequent spread in usual agricultural practice\(^{(6)}\). Genotyping tests to a different drug were sensitive in \(H.\ contortus\) when treatment failure had not yet occurred,\(^{(87)}\) were a similar standard of test developed for IVM resistant genotypes, the early stage of anthelmintic resistance could be detected and even subtle or slow effects could be followed. Cross-sectional studies of \(H.\ contortus\) may reveal variables correlated with IVM resistance and longitudinal research could follow the development of resistance in different flocks. The gold standard of blind randomised controlled trials could measure even slight changes to the \(H.\ contortus\) population genotype. This would contribute to a clear quantitative assessment of competing hypotheses for slowing, limiting or preventing resistance. These potentially useful chemotherapy and management regimes include treating according to clinical symptoms or treating less than 100\% of the sheep with random allocation to treatment or non-treatment groups.
There is a very important advantage of examining genotypes, which is that a warning may be found that a phenotypically sensitive population has the genetic potential in at least some members to be IVM resistant. This pre-clinical resistance would be under selective pressure with traditional whole-flock prophylactic application of this broad-spectrum anthelmintic. Using a molecular test the use of anthelmintics could be adjusted to minimise or eliminate the selective pressure according to best practice at the time. A test that is informative before a problem exists is very helpful in the monitoring of *H. contortus* populations but it may be harder to market this approach. Particularly since the existing tests are not being fully utilised even when there is a need in flocks with suspicions of IVM resistance. Genetic tests will have to be understood to an even greater extent than the existing procedures.

It may be that such a genetic test would first only be able to be used at a specialised remote site with economies of scale. This organisation would be constrained by having to provide results in a time window that is still useful to the farm for making decisions in terms of chemotherapy and management. In terms of treating each animal or portion of a flock this may not be as practical and economically feasible as its use as a flock-wide surveillance method. The barriers to frequent use of a genetic test on many farms are predictable but worth stating. A practical test would have to be robust, sensitive and specific, simple, small, minimally invasive, use the minimum of equipment and not delicate expensive kit, use cheap consumables, be set up and run quickly and be able to be used by minimally trained and qualified personnel.

If such a molecular test could be made widely available then the surveillance and control of IVM resistance in *H. contortus* would be one step closer.
BIBLIOGRAPHY


APPENDIX

TABLE OF CONTENTS

A.1 HAECO_S09_Supercontig_0050550 ≈ CAEEL-PGP-1 87
A.2 HAECO_S09_Supercontig_0045305 ≈ CAEEL-PGP-1 88
A.3 HAECO_S09_Supercontig_0034154 ≈ CAEEL-PGP-1 89
A.4 HAECO_S09_Supercontig_0061621 ≈ CAEEL-PGP-1 90
A.5 HAECO_S09_Supercontig_0006074 ≈ CAEEL-PGP-1 91
A.6 HAECO_S09_Supercontig_0032856 ≈ CAEEL-PGP-1 93
A.7 HAECO_S09_Supercontig_0011236 ≈ CAEEL-PGP-1 94
A.8 HAECO_S09_Supercontig_0044638 ≈ CAEEL-PGP-1 95
A.9 HAECO_S09_Supercontig_0046186 ≈ CAEEL-PGP-1 96
A.10 HAECO_S09_Supercontig_0047549 ≈ CAEEL-PGP-1 97
A.11 HAECO_S09_Supercontig_0024971 ≈ CAEEL-PGP-1 98
A.12 HAECO_S09_Supercontig_0049649 ≈ CAEEL-PGP-2 99
A.13 HAECO_S09_Supercontig_0007774 = HAECO-PGP-2.1.A 100
A.14 HAECO_S09_Supercontig_0038725 = HAECO-PGP-2.1.A 102
A.15 HAECO_S09_Supercontig_0046285 = HAECO-PGP-2.1.A 103
A.16 HAECO_S09_Supercontig_0005977 = HAECO-PGP-2.1.A 104
A.17 HAECO_S09_Supercontig_0001706 ≈ CAEEL-PGP-3 105
A.18 HAECO_S09_Supercontig_0059287 ≈ CAEEL-PGP-3 107
A.19 HAECO_S09_Supercontig_0023983 ≈ CAEEL-PGP-3 108
A.20 HAECO_S09_Supercontig_0000863 ≈ CAEEL-PGP-3/4 109
A.21 HAECO_S09_Supercontig_0001880 ≈ CAEEL-PGP-3/4 110
A.22 HAECO_S09_Supercontig_0057179 ≈ CAEEL-PGP-4 113
A.23 HAECO_S09_Supercontig_0006925 ≈ CAEEL-PGP-9 114
A.24 HAECO_S09_Supercontig_0038098 ≈ CAEEL-PGP-9 119
A.25 HAECO_S09_Supercontig_0046372 ≈ CAEEL-PGP-9 122
A.26 HAECO_S09_Supercontig_0016882 ≈ CAEEL-PGP-9 123
A.27 HAECO_S09_Supercontig_0035472 ≈ CAEEL-PGP-10 124
A.28 HAECO_S09_Supercontig_0025718 ≈ CAEEL-PGP-10 125
A.29 HAECO_S09_Supercontig_0064200 ≈ CAEEL-PGP-10 126
A.30 HAECO_S09_Supercontig_0038139 ≈ CAEEL-PGP-10 127
A.31 HAECO_S09_Supercontig_0037674 ≈ CAEEL-PGP-10 128
A.32 HAECO_S09_Supercontig_0037517 ≈ CAEEL-PGP-10 129
A.33 HAECO_S09_Supercontig_0024002 ≈ CAEEL-PGP-11 130
A.34 HAECO_S09_Supercontig_0024351 ≈ CAEEL-PGP-11 131
A.35 HAECO_S09_Supercontig_0035404 ≈ CAEEL-PGP-11 133
A.36 HAECO_S09_Supercontig_0069412 ≈ CAEEL-PGP-14 134
A.37 HAECO_S09_Supercontig_0004549 ≈ CAEEL-PGP-14 135
A.38 HAECO_S09_Supercontig_0059902 ≈ CAEEL-PGP-13 136
A.39 HAECO_S09_Supercontig_0021658 ≈ CAEEL-PGP-12 137
A.40 HAECO_S09_Supercontig_0050285 ≈ CAEEL-PGP-12 138
A.41 HAECO_S09_Supercontig_0015622 ≈ CAEEL-PGP-14 139
A.42 HAECO_S09_Supercontig_0021166 ≈ CAEEL-PGP-14 140
A.43 HAECO_S09_Supercontig_0010288 ≈ CAEEL-PGP-13 141
A.44 HAECO_S09_Supercontig_0024244 ≈ CAEEL-PGP-14/15 142
A.45 HAECO_S09_Supercontig_0005706 ≈ CAEEL-PGP-15ps 143
A.46 HAECO_S09_Supercontig_0013014 ≈ CAEEL-PGP-1/5/6 145
A.47 HAECO_S09_Supercontig_0055664 ≈ CAEEL-PGP-4/9 146
A.48 HAECO_S09_Supercontig_0055790 ≈ CAEEL-PGP-5/12/14 147
Appendix Sequence Legend:

Green highlights both the exons and introns predicted by the GeneWise algorithm.

Yellow highlights the unknown nucleotides between the contigs within the supercontig and the unknown amino acids due to the identity of the nucleotides being unknown within the supercontig or the GeneWise algorithm being unable to extract coding sequence.

Blue and purple highlight alternative splicing.
**A.2 **

**HAECO_S09_SUPERCONTIG_0045305 = CAEEL-PGP-1**

**Genewise**

**Query protein:** CAEEL-PGP-1

**Target Sequence:** H_S09_0045305

<table>
<thead>
<tr>
<th>E</th>
<th>Query protein</th>
<th>Target Sequence</th>
<th>E0</th>
<th>E0</th>
<th>E0</th>
</tr>
</thead>
</table>
| HAECO_S09_SUPERCONTIG_0045305 | CAEEL-PGP-1 | H_S09_0045305 | 1 | Intron CAG | 88

**Intron CAG**

<table>
<thead>
<tr>
<th>E0</th>
<th>88</th>
<th>170</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAEEL-PGP-1</td>
<td>VPCILVIFIVVVMNSLHFKVFLIQEFKRTNDWATET</td>
<td>VPCILVIFIVVVMNSLHFKVFLIQEFKRTNDWATET</td>
</tr>
<tr>
<td>H_S09_0045305</td>
<td>tttttttgtgatggtatctactaagctaaagtccattatgtgatgatgagtaat</td>
<td>tttttttgtgatggtatctactaagctaaagtccattatgtgatgatgagtaat</td>
</tr>
</tbody>
</table>

**CAEEL-PGP-1**

<table>
<thead>
<tr>
<th>E0</th>
<th>157</th>
</tr>
</thead>
<tbody>
<tr>
<td>VPCILVIFIVVVMNSLHFKVFLIQEFKRTNDWATET</td>
<td>VPCILVIFIVVVMNSLHFKVFLIQEFKRTNDWATET</td>
</tr>
<tr>
<td>H_S09_0045305</td>
<td>gatttagtgcaaacaccatgagatccgatttgaaaagacgaacttcgttttgaatgagtggattacttaaatcgtaaaaggctccattggctgttcgggctatagatgactatattagctatccgagctaac</td>
</tr>
</tbody>
</table>

**CAEEL-PGP-1**

<table>
<thead>
<tr>
<th>E0</th>
<th>201</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>NLERVKEGTGDKIGMAFQYLSQF</td>
</tr>
<tr>
<td>H_S09_0045305</td>
<td>D</td>
</tr>
</tbody>
</table>

**CAEEL-PGP-1**

<table>
<thead>
<tr>
<th>E0</th>
<th>303</th>
</tr>
</thead>
<tbody>
<tr>
<td>g</td>
<td>aagcgaggaggaagcatctaact</td>
</tr>
<tr>
<td>H_S09_0045305</td>
<td>g</td>
</tr>
</tbody>
</table>

**Making a D in phase 2 Intron**

<table>
<thead>
<tr>
<th>E0</th>
<th>305</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAEEL-PGP-1</td>
<td>g</td>
</tr>
<tr>
<td>H_S09_0045305</td>
<td>g</td>
</tr>
</tbody>
</table>

**Intron 1**

<table>
<thead>
<tr>
<th>E0</th>
<th>603</th>
</tr>
</thead>
<tbody>
<tr>
<td>H_S09_0045305</td>
<td>GTGAATT</td>
</tr>
</tbody>
</table>

**//**

**Gene 1 EXONS**

**171 609**

**Exon 171 304 phase 0**

**Exon 533 609 phase 2**

**//**

**Making a D in phase 2 Intron**

<table>
<thead>
<tr>
<th>E0</th>
<th>610</th>
</tr>
</thead>
<tbody>
<tr>
<td>H_S09_0045305</td>
<td>GTGAATT</td>
</tr>
</tbody>
</table>

**//**

**Gene 1 EXONS**

**171 609**

**Exon 171 304 phase 0**

**Exon 533 609 phase 2**

**//**

**Making a D in phase 2 Intron**

<table>
<thead>
<tr>
<th>E0</th>
<th>610</th>
</tr>
</thead>
<tbody>
<tr>
<td>H_S09_0045305</td>
<td>GTGAATT</td>
</tr>
</tbody>
</table>

**//**

**Gene 1 EXONS**

**171 609**

**Exon 171 304 phase 0**

**Exon 533 609 phase 2**

**//**

**Making a D in phase 2 Intron**

<table>
<thead>
<tr>
<th>E0</th>
<th>610</th>
</tr>
</thead>
<tbody>
<tr>
<td>H_S09_0045305</td>
<td>GTGAATT</td>
</tr>
</tbody>
</table>
A.5  
**HAECO_S09_SUPERCONTIG_0006074** = **CAEEL-PGP-1**

gencode

Query protein: CAEEL-PGP-1  CAEEL-PGP-1-R000003P1-1.1:2212AA
Comp Matrix: blosum62.bla
Gap open: 12
Gap extension: 2
Start/End (local)
Start/End (protein): local 0:2212AA
Gene Parei: worm.pf
Codon Table: codon.table
Down: 1e-05
Indel error: 1e-05
Model splice? flat
Model codon bias? tied
Model intron bias? tied
Algorithm 623

K_S00_0006074_4 Introns: CAG

K_S00_0006074_7 Introns: CAG

K_S00_0006074_8 Introns: CAG

K_S00_0006074_9 Introns: CAG

K_S00_0006074_10 Introns: CAG

K_S00_0006074_11 Introns: CAG

K_S00_0006074_12 Introns: CAG

K_S00_0006074_13 Introns: CAG

K_S00_0006074_14 Introns: CAG

K_S00_0006074_15 Introns: CAG

K_S00_0006074_16 Introns: CAG

K_S00_0006074_17 Introns: CAG

K_S00_0006074_18 Introns: CAG

K_S00_0006074_19 Introns: CAG

K_S00_0006074_20 Introns: CAG

K_S00_0006074_21 Introns: CAG

K_S00_0006074_22 Introns: CAG

K_S00_0006074_23 Introns: CAG

K_S00_0006074_24 Introns: CAG

K_S00_0006074_25 Introns: CAG

K_S00_0006074_26 Introns: CAG

K_S00_0006074_27 Introns: CAG

K_S00_0006074_28 Introns: CAG

K_S00_0006074_29 Introns: CAG

K_S00_0006074_30 Introns: CAG

K_S00_0006074_31 Introns: CAG

K_S00_0006074_32 Introns: CAG

K_S00_0006074_33 Introns: CAG

K_S00_0006074_34 Introns: CAG

K_S00_0006074_35 Introns: CAG

K_S00_0006074_36 Introns: CAG

K_S00_0006074_37 Introns: CAG

K_S00_0006074_38 Introns: CAG

K_S00_0006074_39 Introns: CAG

K_S00_0006074_40 Introns: CAG

K_S00_0006074_41 Introns: CAG

K_S00_0006074_42 Introns: CAG

K_S00_0006074_43 Introns: CAG

K_S00_0006074_44 Introns: CAG

K_S00_0006074_45 Introns: CAG

K_S00_0006074_46 Introns: CAG

K_S00_0006074_47 Introns: CAG

K_S00_0006074_48 Introns: CAG

K_S00_0006074_49 Introns: CAG

K_S00_0006074_50 Introns: CAG

K_S00_0006074_51 Introns: CAG

K_S00_0006074_52 Introns: CAG

K_S00_0006074_53 Introns: CAG

K_S00_0006074_54 Introns: CAG

K_S00_0006074_55 Introns: CAG

K_S00_0006074_56 Introns: CAG

K_S00_0006074_57 Introns: CAG

K_S00_0006074_58 Introns: CAG

K_S00_0006074_59 Introns: CAG

K_S00_0006074_60 Introns: CAG

K_S00_0006074_61 Introns: CAG

K_S00_0006074_62 Introns: CAG

K_S00_0006074_63 Introns: CAG

K_S00_0006074_64 Introns: CAG

K_S00_0006074_65 Introns: CAG

K_S00_0006074_66 Introns: CAG

K_S00_0006074_67 Introns: CAG

K_S00_0006074_68 Introns: CAG

K_S00_0006074_69 Introns: CAG

K_S00_0006074_70 Introns: CAG

K_S00_0006074_71 Introns: CAG

K_S00_0006074_72 Introns: CAG

K_S00_0006074_73 Introns: CAG

K_S00_0006074_74 Introns: CAG

K_S00_0006074_75 Introns: CAG

K_S00_0006074_76 Introns: CAG

K_S00_0006074_77 Introns: CAG

K_S00_0006074_78 Introns: CAG

K_S00_0006074_79 Introns: CAG

K_S00_0006074_80 Introns: CAG

K_S00_0006074_81 Introns: CAG

K_S00_0006074_82 Introns: CAG

K_S00_0006074_83 Introns: CAG

K_S00_0006074_84 Introns: CAG
A.6  \textit{HAECO\_S09\_SUPERCONTIG\_0032856 = CAEEL\_PGP\_1}

\begin{verbatim}

geneinfo
Query protein: CAEEL\_PGP\_1
Target Sequence
\_S09\_Supercontig\_0032856
R\_S09\_0032856\_1806
Intron
\_S09\_Supercontig\_0032856
\_S09\_Supercontig\_0032856
CAEEL\_PGP\_1
627
HSGDVQVHDRAEEGQELQYLVATPPQVQFSAAS
665
HSGDVQVHDRAEEGQELQYLVATPPQVQFSAAS
CAEEL\_PGP\_1
627
HSGDVQVHDRAEEGQELQYLVATPPQVQFSAAS

\begin{verbatim}
CAG
<0------[ 856 ]------0>

\begin{verbatim}

\begin{verbatim}

\begin{verbatim}

\begin{verbatim}

\begin{verbatim}

\begin{verbatim}
A.11 \texttt{HAECO\_S09\_SUPERCONTIG\_0024971} = \texttt{CAEEL\_PGP\_1}

\textbf{GeneScan}

\textbf{Query protein:} \texttt{CAEEL-CAEEL-CAEEL-CAEEL}

\textbf{Target Sequence:} 1266

\begin{verbatim}
GRTCIVIAHRFSTVVNADCIAVVKGGVIIEQGTHTELMAKRGFYYELTQKQ
\end{verbatim}

\textbf{CAEEL-CAEEL-CAEEL-CAEEL}

\begin{verbatim}
1297 G THTQLMSEKGAYYKLTQKQ
1297 G THT+LM+++GY+LTQKQ
1297 G THTELMAKRGFYYELTQKQ
\end{verbatim}

\textbf{CAEEL-CAEEL-CAEEL-CAEEL}

\begin{verbatim}
315 MTEKK--1321
315 TVKSASE
\end{verbatim}

\textbf{CAEEL-CAEEL-CAEEL-CAEEL}

\begin{verbatim}
94 gcatagagcctaaggaggtagggagggaagc
94 gcatagagcctaaggaggtagggagggaagc
94 gcatagagcctaaggaggtagggagggaagc
\end{verbatim}

\textbf{CAEEL-CAEEL-CAEEL-CAEEL}

\begin{verbatim}
1117 HTEEH--
1117 TVKEASE
\end{verbatim}

\textbf{CAEEL-CAEEL-CAEEL-CAEEL}

\begin{verbatim}
1117 gcatagagcctaaggaggtagggagggaagc
1117 gcatagagcctaaggaggtagggagggaagc
1117 gcatagagcctaaggaggtagggagggaagc
\end{verbatim}

\textbf{CAEEL-CAEEL-CAEEL-CAEEL}

\begin{verbatim}
1118 
1118 
1118 
\end{verbatim}

\textbf{CAEEL-CAEEL-CAEEL-CAEEL}

\begin{verbatim}
1120 
1120 
1120 
\end{verbatim}

\begin{verbatim}
2997 n [1170 : 4166]
\end{verbatim}

\begin{verbatim}
Making a G in phase 1 intron
\end{verbatim}

\begin{verbatim}
\begin{align*}
\text{EXONS 1 1117} & \text{Exon 1 94 phase 0} \\
\text{Exon 1038 1117 phase 1} & \text{Making a G in phase 1 intron}
\end{align*}
\end{verbatim}
A.13  HAECO_S09_SUPERCONTIG_0007774 = HAECO-PGP-2.1.A

gene

codon start / end position

protein domain

exon

gene structure

 ATG

CTG

CTG

CTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG

TTG
A.14 HAECO_S09_SUPERCONTIG_0038725 = HAECO-PGP-2.1.A

**Target Sequence**

<table>
<thead>
<tr>
<th>Exon</th>
<th>Intron</th>
<th>Gene 1 Exons</th>
</tr>
</thead>
<tbody>
<tr>
<td>372</td>
<td>1208</td>
<td>372 1208</td>
</tr>
</tbody>
</table>

**Gene Information**

<table>
<thead>
<tr>
<th>Gene Symbol</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAECO_AAC38987</td>
<td>1025</td>
<td>1158</td>
</tr>
<tr>
<td>HAECO_PGP-2.1.A</td>
<td>1200</td>
<td>1208</td>
</tr>
</tbody>
</table>

**Query protein**

<table>
<thead>
<tr>
<th>Query</th>
<th>Target</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAECO_AAC38987</td>
<td>HAECO_S09_Supercontig_0038725</td>
</tr>
</tbody>
</table>

**Exon Information**

<table>
<thead>
<tr>
<th>Exon</th>
<th>Phase</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1025</td>
<td>1208</td>
</tr>
</tbody>
</table>

**Intron Information**

<table>
<thead>
<tr>
<th>Intron</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>371</td>
<td>1208</td>
</tr>
</tbody>
</table>

**Target Sequence**

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1025</td>
<td>1158</td>
</tr>
</tbody>
</table>

**Query**

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1025</td>
<td>1158</td>
</tr>
</tbody>
</table>

**Target**

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1025</td>
<td>1158</td>
</tr>
</tbody>
</table>

**Alignments**

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1025</td>
<td>1158</td>
</tr>
</tbody>
</table>

**Alignment Information**

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1025</td>
<td>1158</td>
</tr>
</tbody>
</table>

**Alignment Details**

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1025</td>
<td>1158</td>
</tr>
</tbody>
</table>

**Result**

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1025</td>
<td>1158</td>
</tr>
</tbody>
</table>

**Additional Information**

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1025</td>
<td>1158</td>
</tr>
</tbody>
</table>
query protein: HAECO_AAC38987

target sequence: H_S09_Supercontig_0046285

A.15 HAECO_S09_SUPERCONTIG_0046285 = HAECO-PGP-2.1.A

Gene 1

Making a G in phase 1 intron

Making a G in phase 1 intron

Making a G in phase 1 intron
**A.16 HAECO_S09_SUPERCONTIG_0005977 = HAECO-PGP-2.1.A**

**Query protein:**
HAECO AAC38987 HAECO-PGP-2.1.A AAC38987

**Target Sequence:**
HAECO_S09_Supercontig_0005977

---

**Gene wise**

<table>
<thead>
<tr>
<th>g</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**Genes**

<table>
<thead>
<tr>
<th>Gene</th>
<th>EXONS</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20570 20297</td>
<td>1275 20297</td>
<td></td>
</tr>
</tbody>
</table>

---

**Making a G in phase 1 intron**

```plaintext
Making a G in phase 1 intron
```
A.21 \[\text{HAECO\_S09\_SUPERCONTIG\_0001880} = \text{CAEL\_PGP\_3/4}\]
1143AA
KTVALGPGCGDFTIQLRESQDIPG

AFDEVDARELNLRHLRSQMSLGVQEPIL
STIPSENFCRENIAYGLEQATVDQIENAAKL
ANHHFII

//

>H_S09_0001880__XARCO_S09_Supercontig_0001800__29425-29516bp
_29517-29518bp_Frameshift
_29519-29520bp

GGAAGACGGGATCGTGGGGTTGGACCTCGGTGTGGAAAAAGCACTTCATATTC
ATGGACGATTCTGCGGATCCTG

GCT

TTGGTGAGTGAACGCTCGTGAGACTGAAATTGAGGACCGTCCGTCGAAATGCTC
CTGAC

AGCACTTTCT

AGAACCTATTCTTTCAACTATTCCATCAGAAACTTTTGAGAAAGAAACATT
GGTATCGTCGTGCTGGTAAATGCTAGTTGG CGCAATGCAC

GCCACAACCTCATCA

//
A.23 HAECO_S09_SUPERCONTIG_0006925 = CAEL-PGP-9

geneWISE
Query protein: CAEL-PGP-9
Comp Matrix: P
Gap open: 2
Gap extension: 1
Start/End: 1 644
Sequence: 1 644
Model source: CAEEEL
Model size: 644

target/sequence: 1 644
1 644

CAEL-PGP-9 205
PGP

3166
3752

CAEL-PGP-9 231

3296

CAEL-PGP-9 264

CAEL-PGP-9 247

CAEL-PGP-9 291

CAEL-PGP-9 364

CAEL-PGP-9 351

CAEL-PGP-9 329

CAEL-PGP-9 350

CAEL-PGP-9 403

CAEL-PGP-9 377

CAEL-PGP-9 420

CAEL-PGP-9 383

CAEL-PGP-9 470

CAEL-PGP-9 488

CAEL-PGP-9 517

CAEL-PGP-9 480

CAEL-PGP-9 494

CAEL-PGP-9 512

CAEL-PGP-9 567

CAEL-PGP-9 543

CAEL-PGP-9 503

CAEL-PGP-9 619

CAEL-PGP-9 620

CAEL-PGP-9 611

CAEL-PGP-9 705

CAEL-PGP-9 641

CAEL-PGP-9 733

CAEL-PGP-9 768

CAEL-PGP-9 794

CAEL-PGP-9 822

CAEL-PGP-9 855

CAEL-PGP-9 876

CAEL-PGP-9 890

CAEL-PGP-9 907

CAEL-PGP-9 913

CAEL-PGP-9 972

CAG gtctacacgactcttgcggggagcttatttgaaacaaga

GVKPSNLKGKVTVSNLKFTYPTRPDVPILK

tgtacatctgtttga <1

gttcaggaatctgacctaaaatttttg<0
ggttggacggaagtggtagtgaaaaacttccttacggga

tgcgagtttagggac

gGTATGAT  Intron 2   CAG aaactaagaccgtaactatgtgt

GVSYGFASSIFYFLYASCFRFGLWLIV

GVS +A PG+T+ALVGSSGCGKSTIIQLL R+YNP+ G+

tgcgagtttagggac

GVSYGFASSIFYFLYASCFRFGLWLIV

GVS +A PG+T+ALVGSSGCGKSTIIQLL R+YNP+ G+

tgcgagtttagggac

CAEEL

H_S09_0006925__ 4406

H_S09_0006925__ 3296

H_S09_0006925__ 205

H_S09_0006925__ 231

H_S09_0006925__ 264

H_S09_0006925__ 247

H_S09_0006925__ 291

H_S09_0006925__ 364

H_S09_0006925__ 351

H_S09_0006925__ 329

H_S09_0006925__ 350

H_S09_0006925__ 403

H_S09_0006925__ 377

H_S09_0006925__ 420

H_S09_0006925__ 383

H_S09_0006925__ 470

H_S09_0006925__ 488

H_S09_0006925__ 517

H_S09_0006925__ 480

H_S09_0006925__ 567

H_S09_0006925__ 543

H_S09_0006925__ 503

H_S09_0006925__ 619

H_S09_0006925__ 620

H_S09_0006925__ 705

H_S09_0006925__ 641

H_S09_0006925__ 733

H_S09_0006925__ 768

H_S09_0006925__ 794

H_S09_0006925__ 822

H_S09_0006925__ 855

H_S09_0006925__ 876

H_S09_0006925__ 890

H_S09_0006925__ 907

H_S09_0006925__ 913

H_S09_0006925__ 972
Making a G in phase 1 intron
Making a R in phase 2 intron
Making a E in phase 1 intron
Making a R in phase 1 intron

Exons 2046-9942

Gene 1
Exons 464-942
Exon 2046 2108 phase 8
Exon 2175 2285 phase 8
Exon 2994 3155 phase 9
Exon 3977 3351 phase 13
Exon 3441 3580 phase 17
Exon 2994 3155 phase 12
Exon 4347 4652 phase 16
Exon 4333 4659 phase 16
Exon 4546 5073 phase 11
Exon 5171 5052 phase 11
Exon 5288 5458 phase 13
Exon 5556 5272 phase 15
Exon 7646 7615 phase 16
Exon 7646 7615 phase 16
Exon 8039 8115 phase 14
Exon 8273 8519 phase 2
Exon 8469 8712 phase 3
Exon 9179 9365 phase 12
Exon 9649 9994 phase 12

Making a G in phase 2 intron
Making a N in phase 1 intron
Making a G in phase 1 intron
Making a G in phase 2 intron
Making a N in phase 1 intron

N_C_0006925 KACEO_0006925 Supercontig_0006925

Model splice?        model
Indel error:         1e-06
genewise
Algorithm 623

Model codon bias? flat

Model intron bias? sys

Null model sys

<table>
<thead>
<tr>
<th>H_S09_0006925__20634</th>
</tr>
</thead>
<tbody>
<tr>
<td>9____  253</td>
</tr>
<tr>
<td>9____   95</td>
</tr>
<tr>
<td>9____   75</td>
</tr>
<tr>
<td>PGP</td>
</tr>
<tr>
<td>GVS +A PG+T+ALVG SGCGKSTI+</td>
</tr>
<tr>
<td>GVSLDAQPGQTVALVGSSGCGKSTII</td>
</tr>
<tr>
<td>G++YGFA SI +F YA+ FRFG +LI</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>H_S09_0006925__20580</th>
</tr>
</thead>
<tbody>
<tr>
<td>9____  58</td>
</tr>
<tr>
<td>9____  328</td>
</tr>
<tr>
<td>9____  933</td>
</tr>
<tr>
<td>9____  892</td>
</tr>
<tr>
<td>9____  866</td>
</tr>
<tr>
<td>9____  840</td>
</tr>
</tbody>
</table>

116
A.24  

**HAECO_S09_SUPERCONTIG_0038098 = CAELE-PGP-9**

genecode

<table>
<thead>
<tr>
<th>Query protein:</th>
<th>CAELE-PGP-9</th>
<th>CAELE-PGP-9.071AG01.1.1-12964A</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp Matrix:</td>
<td>12</td>
<td>Global</td>
</tr>
<tr>
<td>Gap extension:</td>
<td>2</td>
<td>Global</td>
</tr>
<tr>
<td>Start/End</td>
<td>[30000000]</td>
<td>CAELE-PGP-9.071AG01.1.1-12964A</td>
</tr>
<tr>
<td>Gene Parei:</td>
<td>Work pf</td>
<td>Codon Table: Codon Table</td>
</tr>
<tr>
<td>Sub error:</td>
<td>1x5</td>
<td>Model codon bias: Flat</td>
</tr>
<tr>
<td>Model intron bias:</td>
<td>91</td>
<td>Model intron bias: Tied</td>
</tr>
<tr>
<td>Algorithm</td>
<td>621</td>
<td></td>
</tr>
</tbody>
</table>

**Model intron bias?**

- tied

**Model codon bias?**

- flat

**Start/End**

- global

**Subs error:**

- 1e

**Gap open:**

- 12

**Comp Matrix:**

- blosum62.bla

**Query protein:**

- CAEEL

**A.24**

**alternative splice**

- intron 13

- $3950 < 6350$

---

**intronic splice**

- intron 13

- $3950 < 6350$

---

**Model intron bias?**

- tied

**Model codon bias?**

- flat

**Subs error:**

- 1e

**Gap open:**

- 12

**Comp Matrix:**

- blosum62.bla

**Query protein:**

- CAEEL

**alternative splice**

- intron 13

- $3950 < 6350$

---

**intronic splice**

- intron 13

- $3950 < 6350$

---

**Model intron bias?**

- tied

**Model codon bias?**

- flat

**Subs error:**

- 1e

**Gap open:**

- 12

**Comp Matrix:**

- blosum62.bla

**Query protein:**

- CAEEL

**alternative splice**

- intron 13

- $3950 < 6350$

---

**intronic splice**

- intron 13

- $3950 < 6350$

---

**Model intron bias?**

- tied

**Model codon bias?**

- flat

**Subs error:**

- 1e

**Gap open:**

- 12

**Comp Matrix:**

- blosum62.bla

**Query protein:**

- CAEEL

**alternative splice**

- intron 13

- $3950 < 6350$

---

**intronic splice**

- intron 13

- $3950 < 6350$

---

**Model intron bias?**

- tied

**Model codon bias?**

- flat

**Subs error:**

- 1e

**Gap open:**

- 12

**Comp Matrix:**

- blosum62.bla

**Query protein:**

- CAEEL

**alternative splice**

- intron 13

- $3950 < 6350$

---

**intronic splice**

- intron 13

- $3950 < 6350$

---

**Model intron bias?**

- tied

**Model codon bias?**

- flat

**Subs error:**

- 1e

**Gap open:**

- 12

**Comp Matrix:**

- blosum62.bla

**Query protein:**

- CAEEL

**alternative splice**

- intron 13

- $3950 < 6350$

---

**intronic splice**

- intron 13

- $3950 < 6350$

---

**Model intron bias?**

- tied

**Model codon bias?**

- flat

**Subs error:**

- 1e

**Gap open:**

- 12

**Comp Matrix:**

- blosum62.bla

**Query protein:**

- CAEEL

**alternative splice**

- intron 13

- $3950 < 6350$

---

**intronic splice**

- intron 13

- $3950 < 6350$

---

**Model intron bias?**

- tied

**Model codon bias?**

- flat

**Subs error:**

- 1e

**Gap open:**

- 12

**Comp Matrix:**

- blosum62.bla

**Query protein:**

- CAEEL

**alternative splice**

- intron 13

- $3950 < 6350$

---

**intronic splice**

- intron 13

- $3950 < 6350$

---

**Model intron bias?**

- tied

**Model codon bias?**

- flat

**Subs error:**

- 1e

**Gap open:**

- 12

**Comp Matrix:**

- blosum62.bla

**Query protein:**

- CAEEL

**alternative splice**

- intron 13

- $3950 < 6350$

---

**intronic splice**

- intron 13

- $3950 < 6350$

---

**Model intron bias?**

- tied

**Model codon bias?**

- flat

**Subs error:**

- 1e

**Gap open:**

- 12

**Comp Matrix:**

- blosum62.bla

**Query protein:**

- CAEEL

**alternative splice**

- intron 13

- $3950 < 6350$

---

**intronic splice**

- intron 13

- $3950 < 6350$

---

**Model intron bias?**

- tied

**Model codon bias?**

- flat

**Subs error:**

- 1e

**Gap open:**

- 12

**Comp Matrix:**

- blosum62.bla

**Query protein:**

- CAEEL

**alternative splice**

- intron 13

- $3950 < 6350$

---

**intronic splice**

- intron 13

- $3950 < 6350$

---

**Model intron bias?**

- tied

**Model codon bias?**

- flat

**Subs error:**

- 1e

**Gap open:**

- 12

**Comp Matrix:**

- blosum62.bla

**Query protein:**

- CAEEL

**alternative splice**

- intron 13

- $3950 < 6350$

---

**intronic splice**

- intron 13

- $3950 < 6350$

---

**Model intron bias?**

- tied

**Model codon bias?**

- flat

**Subs error:**

- 1e

**Gap open:**

- 12

**Comp Matrix:**

- blosum62.bla

**Query protein:**

- CAEEL

**alternative splice**

- intron 13

- $3950 < 6350$

---

**intronic splice**

- intron 13

- $3950 < 6350$

---

**Model intron bias?**

- tied

**Model codon bias?**

- flat

**Subs error:**

- 1e

**Gap open:**

- 12

**Comp Matrix:**

- blosum62.bla

**Query protein:**

- CAEEL

**alternative splice**

- intron 13

- $3950 < 6350$

---

**intronic splice**

- intron 13

- $3950 < 6350$
Query protein: CAEEL-PGP-9

Target Sequence: H_S09_0046372

Exon 366: 455 phase: 1
Exon 1585: 1556 phase: 2

Making a D in phase 2 intron
query protein: CAEEL-PGP-9

target sequence: H_S09_0016882

gene: HAECO_S09_SUPERCONTIG_0016882 = CAEEL-PGP-9

Making a G in phase 1 intron

Intron 1:

Intron 2:

Intron 3:

Gene 1

Exons 98-855

Exon 98 304 phase 0

Exon 369 515 phase 0

Exon 595 721 phase 0

Exon 782 855 phase 1

Making a G in phase 1 intron
A.32  \textbf{HAECO\_S09\_SUPERCONTIG\_0037517} = CAEEL\_PGP\_10

geneInfo
Query protein: CAEEL\_PGP\_10\_CAEEL\_PGP\_10\_C5401\_1\_1-1283AA
compMotor: bisouwmb.bla
Gap open: 12
Gap extension: 2
Start/End: local
Target sequence: _HAECO\_S09\_Supercontig\_0037517_
Gene Parat: work.gf
Codon Table: codon.table
Sub error: 1e-05
Indel error: 1e-05
Model splice? model
Model codon bias? flat
Model intron bias? tied
Null model syn
Algorithm 623

H_S09_0037517\_1

\begin{verbatim}
Intron     CAG<2-----[     : 2126]-2>
10 n [1957 : 1966]
CAEEL\_PGP\_10___  839
GSLLKIAIAPAEERNIDQIFIVAAVYTALIIIKTIFEALGRLFIALYGHGFCSCMRSIMFRK
IMRHGCAYFDEERNSPGRILQRIITDSSTLNKIMESKLDILIPAVICPLFSLAAAMYINWKMALLCSFQFPAYFVIRIVQMKEGTK
\end{verbatim}

Gene 3
EXONS 2128 2916
Exon 2128 2244 phase 0
Exon 2304 2369 phase 0
Exon 2440 2584 phase 0
Exon 2802 2916 phase 1

\begin{verbatim}
Making a L in phase 1 intron
\end{verbatim}

H_S09_0037517\_2

\begin{verbatim}
GTGCGAT  Intron 3   CAGttgtcctttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttttt...
A.33  **HAECO_S09_SUPERCONTIG_0024002 = CAEEL-PGP-11**

---

**Query protein:**  CAEEL-PGP-11  _CAEEL-PGP-11_0001_1_3_1-1270AA

**Target sequence:**  H_S09_0024002  _HAECO_S09_SUPERCONTIG_0024002_1

---

**GENCODE**

**Gene:**

**Exons:**

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start</th>
<th>End</th>
<th>Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5241</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>44</td>
<td>5289</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>109</td>
<td>5424</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>134</td>
<td>5622</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>160</td>
<td>6304</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>188</td>
<td>7259</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>230</td>
<td>7384</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>236</td>
<td>239</td>
<td>0</td>
</tr>
</tbody>
</table>

---

**Translation**

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start</th>
<th>End</th>
<th>Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5241</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>44</td>
<td>5289</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>109</td>
<td>5424</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>134</td>
<td>5622</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>160</td>
<td>6304</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>188</td>
<td>7259</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>230</td>
<td>7384</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>236</td>
<td>239</td>
<td>0</td>
</tr>
</tbody>
</table>

---

**Translation:**

<table>
<thead>
<tr>
<th>Exon</th>
<th>Start</th>
<th>End</th>
<th>Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>5241</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>44</td>
<td>5289</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>109</td>
<td>5424</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>134</td>
<td>5622</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>160</td>
<td>6304</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>188</td>
<td>7259</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>230</td>
<td>7384</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>236</td>
<td>239</td>
<td>0</td>
</tr>
</tbody>
</table>
**A.34**

**HAECO_S09SUPERCONTIG_0024351 = CAEL-PGP-11**

**genebase**

**Query protein:** CAEL-PGP-11  CAEL-PGP-11 DROIDs 3 1:1270A

**Target Sequence:** CAEEL_S09SUPERCONTIG_0024351

---

**CAEL-PGP-11**

**Gene:**

**Gene Symbol:** CAEEL

**Gene Name:** H_S09_0024351

**Exon Numbers:**

- Exon 1: 2932-2931
- Exon 2: 3257-3218
- Exon 3: 5092-4987
- Exon 4: 7522-7486
- Exon 5: 8752-8635
- Exon 6: 9791-9636
- Exon 7: 10961-10790

**Intron Numbers:**

- Intron 1: 2932-2931
- Intron 2: 3257-3218
- Intron 3: 5092-4987
- Intron 4: 7522-7486
- Intron 5: 8752-8635
- Intron 6: 9791-9636
- Intron 7: 10961-10790

**Protein:**

**Protein Name:** CAEEL

**Protein Accession:** H_S09_0024351

**Protein Symbol:** CAEEL

---

**Alignment:**

**Acceptor Site:** GT

**Donor Site:** AC

**Start Codon:** ATG

**End Codon:** TAA

**ORF Length:** 1773

**AA Length:** 591

**Intron Positions:**

- Intron 1: 2932-2931
- Intron 2: 3257-3218
- Intron 3: 5092-4987
- Intron 4: 7522-7486
- Intron 5: 8752-8635
- Intron 6: 9791-9636
- Intron 7: 10961-10790

**Coding Strand:** 1

**Coding Frame:** 2

**Start:** 1

**End:** 591

**ORF Sequence:**

```
<table>
<thead>
<tr>
<th>Exon 1</th>
<th>Exon 2</th>
<th>Exon 3</th>
<th>Exon 4</th>
<th>Exon 5</th>
<th>Exon 6</th>
<th>Exon 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>2932-2931</td>
<td>3257-3218</td>
<td>5092-4987</td>
<td>7522-7486</td>
<td>8752-8635</td>
<td>9791-9636</td>
<td>10961-10790</td>
</tr>
</tbody>
</table>
```

**Translation:**

```
ALAKDVYFSYSTRNQPVLNLIGSCEAQIEFRDVRFKYFTRDNEVLQP
```

**Amino Acid Sequence:**

```
ALAKDVYFSYSTRNQPVLNLIGSCEAQIEFRDVRFKYFTRDNEVLQP
```

---

**Alternative Splice:**

```
//
```

---

**Comparison:**

```
//
```

---

**Analysis:**

```
//
```

---

**Conclusion:**

```
//
```

---

**References:**

```
//
```

---

**Table:**

```
//
```

---

**Figure:**

```
//
```

---

**Graph:**

```
//
```

---

**Diagram:**

```
//
```
### Genes

**Query protein:** CAEEL-PGP-12

| Target Sequence | CAEEL-PGP-12 | H_S09_0021658
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>H_S09_0021658</td>
<td>1</td>
<td>Intron</td>
</tr>
<tr>
<td>H_S09_0021658</td>
<td>1</td>
<td>Intron</td>
</tr>
<tr>
<td>Genewise</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genewise</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Target Sequence**

<table>
<thead>
<tr>
<th>H_S09_0021658</th>
<th>1</th>
<th>Target Sequence</th>
<th>CAEEL-PGP-12</th>
</tr>
</thead>
<tbody>
<tr>
<td>H_S09_0021658</td>
<td>1</td>
<td>Intron</td>
<td>CAG</td>
</tr>
<tr>
<td>H_S09_0021658</td>
<td>1</td>
<td>Intron</td>
<td>CAG</td>
</tr>
</tbody>
</table>

### Exons

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Exon 375 1273: exon 375 1273 phase 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon 1063 1158: exon 1063 1158 phase 0</td>
<td></td>
</tr>
<tr>
<td>Exon 385 503: exon 385 503 phase 0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gene 2</th>
<th>Exon 375 1273: exon 375 1273 phase 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon 126 1273: exon 126 1273 phase 0</td>
<td></td>
</tr>
</tbody>
</table>

---

**A.39** **HAECO_S09_SUPERCONTIG_0021658 = CAEEL-PGP-12**

---

**Gene:**

- **Query protein:** CAEEL-PGP-12
- **Target Sequence:** H_S09_0021658

**Exons**

- **Gene 1:** Exon 375 1273: exon 375 1273 phase 0
- **Gene 2:** Exon 375 1273: exon 375 1273 phase 0

---

**Genewise**

- **Query protein:** CAEEL-PGP-12
- **Target Sequence:** H_S09_0021658

---

**Genewise**

- **Query protein:** CAEEL-PGP-12
- **Target Sequence:** H_S09_0021658

---

**Genewise**

- **Query protein:** CAEEL-PGP-12
- **Target Sequence:** H_S09_0021658

---

**Genewise**

- **Query protein:** CAEEL-PGP-12
- **Target Sequence:** H_S09_0021658
A.40  **HAECO_S09_SUPERCONTIG_0050285 = CAEEL-PGP-12**

**genesis**

**Query protein:**  
CAEEL-PGP-12  
CAEEL-PGP-12

**Target Sequence:**  
H_S09_0050285  
H_S09_0050285

**Exons:**

- **60-420:**  
  EXONS 60 420  
  Exon 60 148 phase 0  
  Exon 30 6 420 phase 2  
  Making a R in phase 2: intron

- **593-660:**  
  H_S09_0050285___HAECO_S09_Supercontig_0050285  
  _60—420bp_AA  
  _CAEEL  
  _PGP  
  _12_  
  593—660AA

**Alignment:**

- **60-420:**  
  XTAEVSVEVAVDPPKQ  
  CTSSVPHQKVQDDRQLTTRDQA

- **593-660:**  
  YQALPAQKLYANQRTIMESRHSTLIGL
  YQALPAQKLYANQRTIMESRHSTLIGL

**Intron:**

- **60-420:**  
  Intron 1  
  CAGatatagttcaggcacattatag  
  <2—149 : 305—2>gaactataaaagtatctaccagca

- **593-660:**  
  CAGatatagttcaggcacattatag  
  <2—149 : 305—2>gaactataaaagtatctaccagca

**Gene:**

- **3:**  
  Exon 60 420  
  Exon 60 148 phase 0  
  Exon 30 6 420 phase 2  

- **593-660:**  
  H_S09_0050285___HAECO_S09_Supercontig_0050285  
  _60—420bp_AA  
  _CAEEL  
  _PGP  
  _12_  
  593—660AA

**Untranslated Regions:**

- **593-660:**  
  YQALPAQKLYANQRTIMESRHSTLIGL
  YQALPAQKLYANQRTIMESRHSTLIGL
**A.44**  
**HAECO_SO9_SUPERCONTIG_0024244 = CAEL-PGP-14/15**

### genecolos

**Query protein:** CAEL-PGP-14  
**Comp Matrix:** biseq.1.bk  
**Gap open:** 12  
**Gap extension:** 2  
**Start/End sequence:** local  
**Target sequence:** CAEEL_H_S09_0024244  
**Gene Parsys:** work.df  
**Codon Table:** codon.table  
**Sub error:** 1e-05  
**Model explicit:** model  
**Model codon bias:** flat  
**Model intron bias:** tied  
**Model splice:** model  
**Start/End (protein):** local  
**Target Sequence:** CAEEL_H_S09_0024244___  
**Algorithm:** 621  

---

**R_009_0024244**  
**Intron TAG**  

**CAEEL-PGP-14_701**  
**Gene Parsys**  
**Target sequence** CAEEL_H_S09_0024244  
**Start/End sequence** local  
**Gene Parsys** work.df  
**Codon Table** codon.table  
**Sub error** 1e-05  
**Model explicit** model  
**Model codon bias** flat  
**Model intron bias** tied  
**Model splice** model  
**Start/End (protein)** local  
**Target Sequence** CAEEL_H_S09_0024244___  
**Algorithm** 611  

---

**R_009_0024244**  
**Intron TAG**  

**CAEEL-PGP-14_744**  
**Gene Parsys**  
**Target sequence** CAEEL_H_S09_0024244  
**Start/End sequence** local  
**Gene Parsys** work.df  
**Codon Table** codon.table  
**Sub error** 1e-05  
**Model explicit** model  
**Model codon bias** flat  
**Model intron bias** tied  
**Model splice** model  
**Start/End (protein)** local  
**Target Sequence** CAEEL_H_S09_0024244___  
**Algorithm** 611  

---

**R_009_0024244**  
**Intron TAG**  

**CAEEL-PGP-14_770**  
**Gene Parsys**  
**Target sequence** CAEEL_H_S09_0024244  
**Start/End sequence** local  
**Gene Parsys** work.df  
**Codon Table** codon.table  
**Sub error** 1e-05  
**Model explicit** model  
**Model codon bias** flat  
**Model intron bias** tied  
**Model splice** model  
**Start/End (protein)** local  
**Target Sequence** CAEEL_H_S09_0024244___  
**Algorithm** 611  

---

**R_009_0024244**  
**Intron TAG**  

**CAEEL-PGP-14_826**  
**Gene Parsys**  
**Target sequence** CAEEL_H_S09_0024244  
**Start/End sequence** local  
**Gene Parsys** work.df  
**Codon Table** codon.table  
**Sub error** 1e-05  
**Model explicit** model  
**Model codon bias** flat  
**Model intron bias** tied  
**Model splice** model  
**Start/End (protein)** local  
**Target Sequence** CAEEL_H_S09_0024244___  
**Algorithm** 611  

---

**R_009_0024244**  
**Intron TAG**  

**CAEEL-PGP-14_854**  
**Gene Parsys**  
**Target sequence** CAEEL_H_S09_0024244  
**Start/End sequence** local  
**Gene Parsys** work.df  
**Codon Table** codon.table  
**Sub error** 1e-05  
**Model explicit** model  
**Model codon bias** flat  
**Model intron bias** tied  
**Model splice** model  
**Start/End (protein)** local  
**Target Sequence** CAEEL_H_S09_0024244___  
**Algorithm** 611  

---

**R_009_0024244**  
**Intron TAG**  

**CAEEL-PGP-14_882**  
**Gene Parsys**  
**Target sequence** CAEEL_H_S09_0024244  
**Start/End sequence** local  
**Gene Parsys** work.df  
**Codon Table** codon.table  
**Sub error** 1e-05  
**Model explicit** model  
**Model codon bias** flat  
**Model intron bias** tied  
**Model splice** model  
**Start/End (protein)** local  
**Target Sequence** CAEEL_H_S09_0024244___  
**Algorithm** 611  

---

**R_009_0024244**  
**Intron TAG**  

**CAEEL-PGP-14_911**  
**Gene Parsys**  
**Target sequence** CAEEL_H_S09_0024244  
**Start/End sequence** local  
**Gene Parsys** work.df  
**Codon Table** codon.table  
**Sub error** 1e-05  
**Model explicit** model  
**Model codon bias** flat  
**Model intron bias** tied  
**Model splice** model  
**Start/End (protein)** local  
**Target Sequence** CAEEL_H_S09_0024244___  
**Algorithm** 611  

---

**R_009_0024244**  
**Intron TAG**  

**CAEEL-PGP-14_937**  
**Gene Parsys**  
**Target sequence** CAEEL_H_S09_0024244  
**Start/End sequence** local  
**Gene Parsys** work.df  
**Codon Table** codon.table  
**Sub error** 1e-05  
**Model explicit** model  
**Model codon bias** flat  
**Model intron bias** tied  
**Model splice** model  
**Start/End (protein)** local  
**Target Sequence** CAEEL_H_S09_0024244___  
**Algorithm** 611  

---

**R_009_0024244**  
**Intron TAG**  

**CAEEL-PGP-14_950**  
**Gene Parsys**  
**Target sequence** CAEEL_H_S09_0024244  
**Start/End sequence** local  
**Gene Parsys** work.df  
**Codon Table** codon.table  
**Sub error** 1e-05  
**Model explicit** model  
**Model codon bias** flat  
**Model intron bias** tied  
**Model splice** model  
**Start/End (protein)** local  
**Target Sequence** CAEEL_H_S09_0024244___  
**Algorithm** 611  

---

**R_009_0024244**  
**Intron TAG**  

**CAEEL-PGP-14_1000**  
**Gene Parsys**  
**Target sequence** CAEEL_H_S09_0024244  
**Start/End sequence** local  
**Gene Parsys** work.df  
**Codon Table** codon.table  
**Sub error** 1e-05  
**Model explicit** model  
**Model codon bias** flat  
**Model intron bias** tied  
**Model splice** model  
**Start/End (protein)** local  
**Target Sequence** CAEEL_H_S09_0024244___  
**Algorithm** 611  

---

**R_009_0024244**  
**Intron TAG**  

**CAEEL-PGP-14_1050**  
**Gene Parsys**  
**Target sequence** CAEEL_H_S09_0024244  
**Start/End sequence** local  
**Gene Parsys** work.df  
**Codon Table** codon.table  
**Sub error** 1e-05  
**Model explicit** model  
**Model codon bias** flat  
**Model intron bias** tied  
**Model splice** model  
**Start/End (protein)** local  
**Target Sequence** CAEEL_H_S09_0024244___  
**Algorithm** 611  

---
A.45  **HAECO_S09_SUPERCONTIG_0005706 = CAEEL-PGP-15PS**

### Gene 1: CAEEL-PGP-15PS

**GeneSeq**
- Query protein: CAEEL-PGP-15PS (F2210_4_not_counting_X_1-1220AA)
- Query Matrix: F2210_4_not_counting_X_1-1220AA
- Gap open: 12
- Gap extension: 2
- Start/End: local
- Target sequence: CAEEL-PGP-15PS (F2210_4_not_counting_X_1-1220AA)
- Start/End (proteins): local
- Gene Name: CAEEL-PGP-15PS
- Model codon bias: flat
- Model intron bias:
- Aligner: ClustalW
- Algorithm: 631

**Exon Information**
- **Exon 8758 8871 phase 0**
- **Exon 4483 4**
- **Exon 4342 4427 phase 0**
- **Exon 57 237 phase 0**

**Sequence Alignment**

```
<table>
<thead>
<tr>
<th>Gene: 09</th>
<th>Gene: 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>1170</td>
<td>1170</td>
</tr>
<tr>
<td>585</td>
<td>585</td>
</tr>
<tr>
<td>951</td>
<td></td>
</tr>
<tr>
<td>1292</td>
<td></td>
</tr>
<tr>
<td>1361</td>
<td></td>
</tr>
<tr>
<td>534</td>
<td></td>
</tr>
<tr>
<td>1170</td>
<td></td>
</tr>
<tr>
<td>279</td>
<td></td>
</tr>
<tr>
<td>117</td>
<td></td>
</tr>
<tr>
<td>854</td>
<td></td>
</tr>
<tr>
<td>917</td>
<td></td>
</tr>
</tbody>
</table>
```

**Model Information**
- **Subs error:** 1e-04
- **Model:** pgp
- **Intron bias:** tied

**Model Output**
- **Model intron bias:** tied
- **Model codon bias:** flat
- **Model error:** 1e-05
- **Model extension:** 2
- **Start/End:** local

**Match Information**
- **Gene Name:** CAEEL-PGP-15PS
- **Aligner:** ClustalW
- **Algorithm:** 631

**Exon Information**
- **Exon 8758 8871 phase 0**
- **Exon 4483 4**
- **Exon 4342 4427 phase 0**
- **Exon 57 237 phase 0**

**Sequence Alignment**

```
<table>
<thead>
<tr>
<th>Gene: 09</th>
<th>Gene: 15</th>
</tr>
</thead>
<tbody>
<tr>
<td>55</td>
<td>55</td>
</tr>
<tr>
<td>1170</td>
<td>1170</td>
</tr>
<tr>
<td>585</td>
<td>585</td>
</tr>
<tr>
<td>951</td>
<td></td>
</tr>
</tbody>
</table>
```

**Model Information**
- **Subs error:** 1e-04
- **Model:** pgp
- **Intron bias:** tied

**Model Output**
- **Model intron bias:** tied
- **Model codon bias:** flat
- **Model error:** 1e-05
- **Model extension:** 2
- **Start/End:** local

**Match Information**
- **Gene Name:** CAEEL-PGP-15PS
- **Aligner:** ClustalW
- **Algorithm:** 631
H_2_013014_S09_0013014

CAEEL

H_2_013014_S09_0013014

CAEEL

H_2_013014_S09_0013014

CAEEL

H_2_013014_S09_0013014

CAEEL

H_2_013014_S09_0013014

genewise

A.46

HAECO_S09_SUPERCONTIG_0013014 = CAEEL- PGP-1/5/6
A.47  \textbf{HAECO\_S09\_SUPERCONTIG\_0055664 = CAEEL\-PGP\-4/9}

geneinfo

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
\textbf{Gene} & \textbf{Target Sequence} & \textbf{Query protein} & \textbf{Target Sequence} \\
\hline
\textbf{H\_S09\_0055664} & 1 & Intron & CAG \\
\hline
\textbf{CAEEL\-PGP\-9} & 733 & VFNPSKQHMQMHRQHRLNHEVYLFAGQVLOGLPGCLPFAAE \textbf{HAECO\_S09\_SUPERCONTIG\_0055664} & 733 \\
\hline
\textbf{H\_S09\_0055664} & 113 & H\textsc{ttatg}\textsc{gagtatgatcttcttg} & 112 \\
\hline
\textbf{CAEEL\-PGP\-9} & 779 & DTTVEKVIYV & 778 \\
\hline
\textbf{H\_S09\_0055664} & 251 & G\textsc{gtagcttac} & 250 \\
\hline
\end{tabular}
\end{table}

\begin{align*}
\text{CAEEL\-PGP\-9} & \quad \text{H\_S09\_0055664} \\
& \quad \text{CAEEL\-PGP\-9} \\
& \quad \text{H\_S09\_0055664}
\end{align*}

\begin{align*}
& \quad \text{H\_S09\_0055664} \\
& \quad \text{H\_S09\_0055664}
\end{align*}
A.48  **HAECO_S09_SUPERCONTIG_0055790 = CAEEL-PGP-5/12/14**

**Query protein:** CAEEL-PGP-14  |  **Target Sequence:** H_S09_0055790  |  **Target Sequence:** HAECO_S09_Supercontig_0055790

**GeneInfo**

```
Gene: 1
Exons: 586-469
Exon 586-469 phase 0
```

**Target Sequence**

```
H_S09_0055790___HAECO_S09_Supercontig_0055790

H_S09_0055790___-2162
CAG<--|

[     :  587]
|

-0>

10 n [1057 : 1048]
```

```
CAEEL-POP-14___608 GALAANGGITITIAHALISTITIHYPEROTYEE

AIH A+HTITI+IALISTI+ K+ E+K+
VALAAGGITITIIHALISTIHYPERO

H_S09_0055790__-588 qptagqqqagaqagqaaegqaaegqaaegqaaegqaaegqaaegqaaegq

ttaagqagqagqagqaaegqaaegqaaegqaaegqaaegqaaegqaaegqaaegqaaegq
```

```
H_S09_0055790__668 OTAGAG 4tron 1

<-------------------------1-1->
```

```
// Gene: 3
Exons: 586-469
Exon 586-469 phase 0
```

```
H_S09_0055790__HAECO_S09_Supercontig_0055790

H_S09_0055790__-586

GTAAGAG  Intron

<---|

[468  : 468]
```

```
CAEEL-POP-14___608 GALAANGGITITIIHALISTITIHYPERO

VAHLAAGGITITIIHALISTIHYPERO

H_S09_0055790__-469

ggtagggagaaaagagcctaaacggaaatgagagagggtg
```

```
H_S09_0055790__-468

tctaaaccaggcctttcagtgctgataatattaagattac
```

```
H_S09_0055790__-467

gtgtaggaaaacccaattgtttttagaatcgagaaatga
```

```
// Gene: 4
Exons: 586-469
```

```
H_S09_0055790__HAECO_S09_Supercontig_0055790

H_S09_0055790__-585

gttgctttgaatgaagcggcgaaaggaagaacaaccatcgtcatagcatacgtcataggagtactattcgtgatgtaaagaaaatatatgtcatggaaaagggaaaagtagttgag
```

```
H_S09_0055790__-584

tcag
```

```
//