50HGP Compara families analysis

We built a custom EnsemblCompara (Flicek et al. 2013) database for the 50 Helminth genome initiative. This includes 91 species (56 Nematodes [4 free-living], 25 Platyhelminthes [1 free-living], 10 outgroups species). From a combined dataset of 1,640,841 genes (all genes in the 91 species), a total of 109,571 gene trees, based on sequence similarity, were created by the EnsemblCompara pipeline, using several tree-inference methods. Most of these families have sizes between 2 and 5 genes (Figure 1), however 1655 families have more than 100 genes, with the largest family containing 1219 genes.

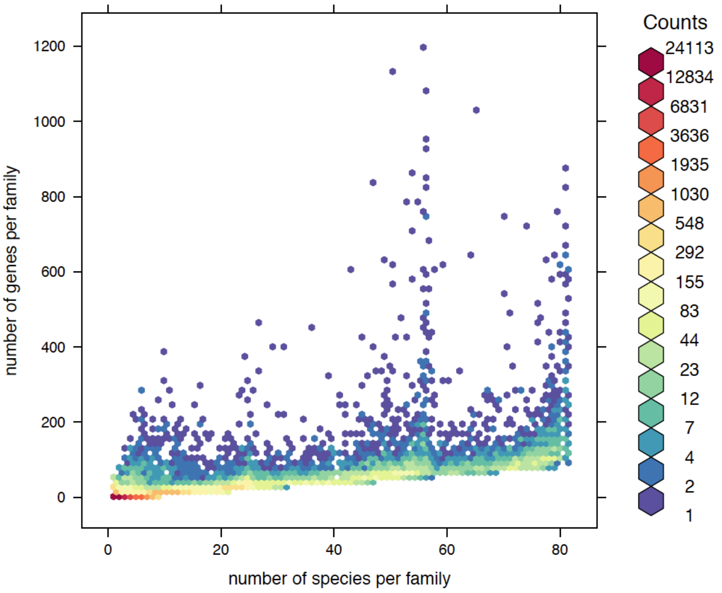


Figure 1. 2D histogram distribution of all 50HG Compara gene families (without any filtering).

While our 50HG gene annotations went through an early step of filtering out transposable-element-related genes, this was not done extensively for some species, therefore we added post-filtering step for gene families where the majority of genes contain transposon-related pfam domains. This excluded 1223 (updated 21-May) gene families from further analysis.

For our preliminary analysis we decided to use a subset of 33 species (instead of the 91 species) for which we have more complete genome assemblies and more robust gene annotations, this includes 4 nematode free-living species and excludes all outgroup species. These 33 species subset has exemplary species from the all the nematode and platyhelminth clades and classes under study. Moreover, as small gene families are not highly informative, we decided to exclude from analysis gene families with less than 6 genes. This smaller dataset for the 33 species (also including transposon filtering) provides us 357,053 genes in 18,436 families for further analysis (Figure 2). The median and mean family size in this dataset is 12 and 19.23 genes, respectively.

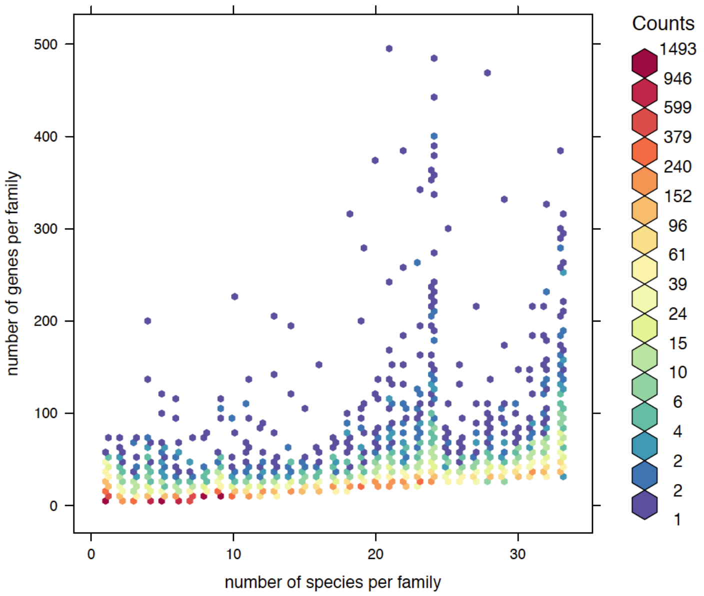


Figure 2. 2D histogram of family size distribution for the subset of 33 species.

Our preliminary analysis is based on these topics:

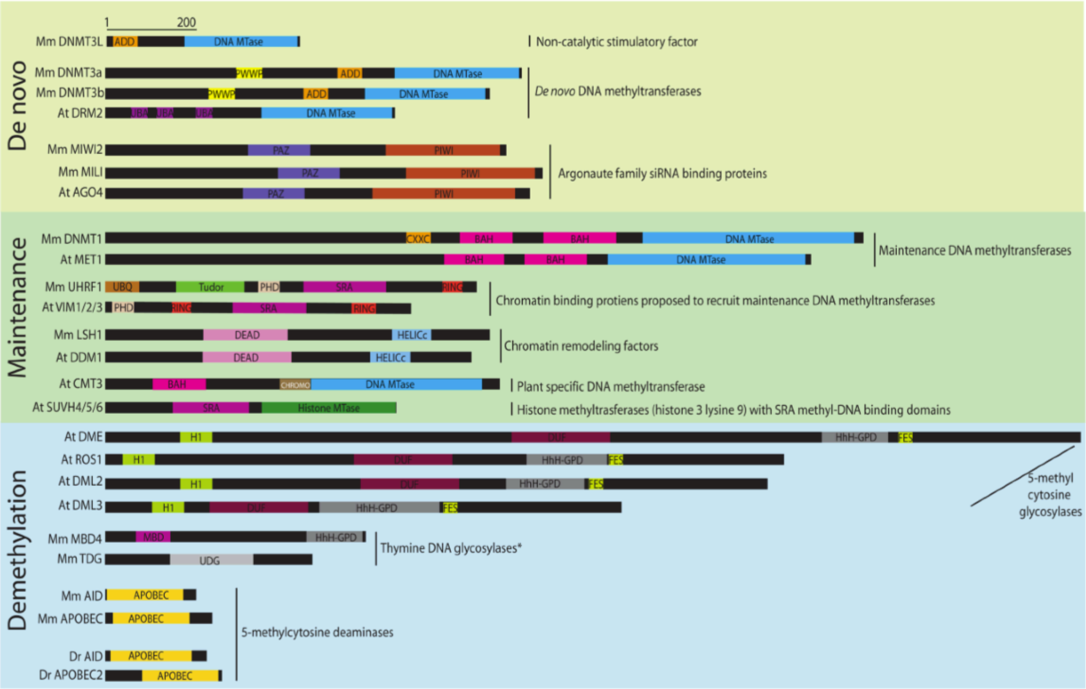
1) Presence of DNA methylation machinery
2) Presence of RNAi-related genes
3) Data mining for gene expansions and families potentially involved in parasitism
4) Exploring families potentially involved in parasitism

Presence of DNA methylation machinery

Methylation of position 5 of cytosines on nuclear DNA has been found in every vertebrate examined, as well as in certain fungi, plant and invertebrate species, including some insect species (although at low prevalence), certain nematodes and potentially all platyhelminthes (Antequera, Tamame, and Villanuevaz 1984; Gao, Wang, and Liu 2014; Hu et al. 2015; Law and Jacobsen 2011; Raddatz et al. 2013).

The DNA methylation rates are very different between organisms, and life-stages. Cytosine DNA methylation often occurs in CpG dinucleotides, impacting gene expression mostly through a close relationship with methyl-CpG-binding domain proteins (MDB proteins) and subsequent recruitment of histone-related proteins that eventually lead to formation of compact inactive chromatin (Roloff, Ropers, and Nuber 2003).

The addition of methyl groups to cytosines is performed by DNA methyltransferases (DNMTs). In mammals, 5 candidate DNMTs exist, DNMT1, DNMT2/TRDMT1, DNMT3A, DNMT3B, DNMT3L (see Box 1) (Law and Jacobsen 2011; Manuscript 2013). DNMT1 is associated with maintenance of existing methylation sites after DNA replication, while DNMT3(B/L) is responsible for *de novo* DNA methylation. DNMT2 in vertebrates shows to have more affinity for aspartic acid tRNA substrate rather than DNA (Goll et al. 2006), however, in some invertebrate and unicellular eukaryote species this protein works as a (low efficiency) DNA methyltransferase, and is in fact the only candidate protein for cytosine DNA methylation in several species (Raddatz et al. 2013; Schaefer and Lyko 2010). In Platyhelminthes DNMT2 seems to be active on DNA (Geyer et al. 2013) while in nematodes sequence analysis relates *dnmt2* genes to tRNA methyltransferase genes (Gao et al. 2012).



Box 1. From (Law and Jacobsen 2011).

DNA methylation has been extensively studied, however, it has been controversial for being missing (or with very low prevalence) from several model organisms including *Saccharomyces cerevisiae, Drosophila melanogaster* and *C. elegans*, but present in other related non-model species of fungi, nematodes and platyhelminths, and in fact there is some astonishment as to how a mechanism that is essential in so many organisms seems to have been lost multiple times throughout Metazoan evolution (Capuano et al. 2014; Gao, Wang, and Liu 2014; Gao et al. 2012; Geyer et al. 2013).

In 11 surveyed nematode species (from genus: Ascaris, Strongyloides, Melodogyne, Bursaphelenchus, Brugia, Pristionchus, Caenorhabditis) (from clades I, III, IV and V), only *Trichinella spiralis* (clade I) possessed a complete set of DNA methylation machinery: DNMT1, DNMT2 and DNMT3 (Gao, Wang, and Liu 2014; Gao et al. 2012). Also, so far *T. spiralis* has been the only nematode where DNA methylation has been clearly demonstrated at non-negligible levels (Gao, Wang, and Liu 2014; Hu et al. 2015). Most nematode species retained the *dnmt1* gene, which seems not to be able to generate *de novo* methylation (at least in *C. elegans*), while a few other nematode species (including *Ascaris suum*) also contain the enigmatic *dnmt2* gene, which shows unknown or reduced DNA methylation activity in nematodes. Also, some MBD genes have been described for *T. spiralis*, without orthologs in other nematodes (Gao, Wang, and Liu 2014).

In Platyhelminthes, even tough initially controversial (Raddatz et al. 2013), DNA methylation has been demonstrated in *Schistosoma mansoni*, *Echinoccocus multilocularis*, *Protopolystoma xenopodis*, *Fasciola hepatica* and *Polycelis nigra,* which comprise all 4 Platyhelminth classes (Geyer et al. 2013). Interestingly, in all these species the only DNA methyltransferase candidate is a *dnmt2*-like gene, and its expression has been show to be correlated with methylation levels. Also, human-like *mdb2/3* genes have been found in *S. mansoni*, suggesting presence of epigenetic control from DNA methylation as in vertebrate species, although the inheritance of DNA methylation has not been described yet.

With the aim of exploring the presence of DNA methylation machinery on all our 50HGP species, we set to look for ortholog genes (from EnsemblCompara families) of the literature-described DNMTs and MBD genes in *Trichinella spiralis*, *Schistosoma* *mansoni* and *Homo sapiens* (Gao, Wang, and Liu 2014; Gao et al. 2012; Geyer et al. 2013; Law and Jacobsen 2011; Roloff, Ropers, and Nuber 2003) (Genes: T. spiralis: EFV54759.1,EFV58204.1,EFV60295.1,EFV62390,EFV60964,EFV57780; S. mansoni: NCBI accessions: HM991456, HM991455; H. sapiens: ENSG00000130816,ENSG00000107614,ENSG00000119772,ENSG00000142182,ENSG00000169057,ENSG00000141644,ENSG00000134046,ENSG00000071655,ENSG00000129071,ENSG00000076108,ENSG00000123636,ENSG00000136169,ENSG00000143379). This search was also complemented by a search of Compara families containing Pfam domains that are prevalent in the described DNMTs and MBDs genes (searching for DNA Methyltransferases by the Pfam domain (PF00145) 'C-5 cytosine-specific DNA methylase'; searching for MBD proteins, keeping Compara families with any combination between, 'Methyl-CpG binding domain' and one or more of the following: 'SET domain', 'Pre-SET motif', 'Bromodomain', 'PHD-finger').

The search for DNMTs and MBDs over our 33 high-quality helminth species set returned a total of 148 genes in 11 families, 53 genes being DNMTs (Figure 1). For the full set of 91 species this harbored 470 genes (151 DNMTs) also in 11 families (Figure 2). However, the presence or absence of a gene in any species outside the selected 33 should not be considered hard evidence (that is the reason both figures are provided).

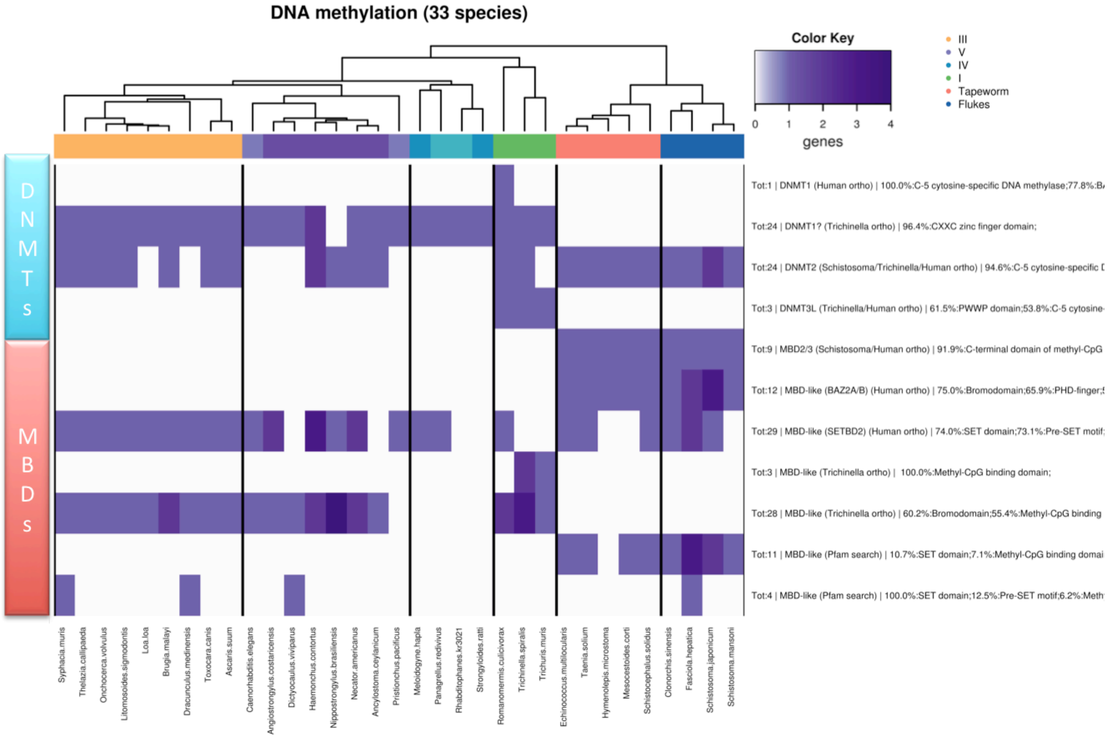


Figure 1. Presence of DNA methylation-related genes over 33 helminth species. Compara families selected from orthologs of *T. spiralis, S. mansoni* and *H. sapiens* DNMT and MBD genes and from presence of methylation-related Pfam domains.

Human *dnmt1* (first row in the figures) does not contain helminth orthologs except for a single *Romanomermis culcivorax* gene. Presence of DNA methylation-related genes non-orthologous to *C. elegans* in *R. culcivorax* has been suggested (Schiffer et al. 2013). The Compara tree building clustered this gene together with vertebrate and insect genes (Figure 2), however, as this is the only helminth with a gene in this cluster, the most likely explanation for is an artifactual clustering, as the multi-alignment shows *R. culcivorax* protein to be extremely diverse compared to its supposed orthologs.

No ortholog genes in the family containing the *T. spiralis dnmt1* (second row in the figures) enclose a methyltransferase Pfam domain, however, such genes are present in all high-quality nematode genomes except *Nippostrongylus brasiliensis* (Figure 1).

Note that this gene is present in *C. elegans*, which is shown to have very low methylation levels, therefore the presence of this single gene in nematodes is not indicative of a functional DNA methylation system.

Of note, the described *dnmt3* gene in *T. spiralis* shows to be present in all Clade I nematode species (Figure 2), and is exclusive to Clade I, as reported by Gao *et al.* 2012. This gene is supposed to encode the DNA methyltransferase responsible for *de novo* methylation in nematodes, and therefore the key for a DNA-methylation-active nematode, therefore these results, combined with an inability to find alternative families of DNA methyltransferase genes using Pfam domain searches, suggests that functional DNA methylation machinery may only be present in Clade I nematodes.

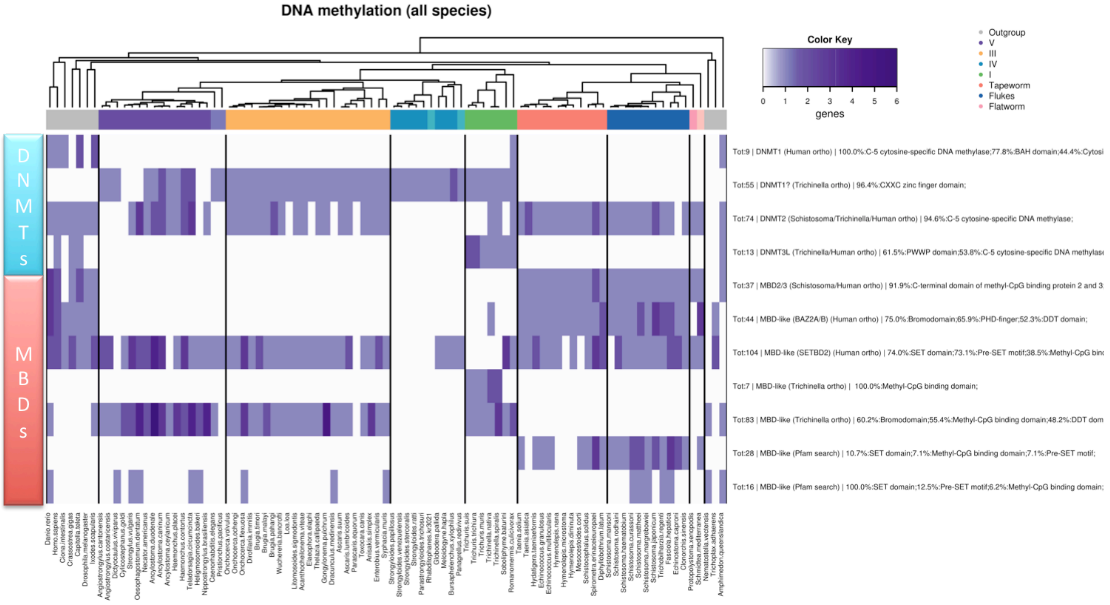


Figure 2. Presence of DNA methylation-related genes over all 91 50HGP species. Compara families selected from orthologs of *T. spiralis, S. mansoni* and *H. sapiens* DNMT and MBD genes and from presence of methylation-related Pfam domains.

In nematodes, *dnmt2* genes seem to have been lost several times independently (including in *C. elegans* and several other Clade V nematodes and lost in all Clade IV nematodes), which again supports the notion that the DNMT2 protein has limited activity in nematodes. Note that we find dnmt2 genes present in more clades than shown in Gao *et al*. 2012. Conversely, in Platyhelminthes, *dnmt2* genes (where its DNA methyltransferase activity has been shown) are conserved in all 9 surveyed species, confirming Geyer *et al.* 2013 results on the conservation of this gene across the Platyhelminthes phylum. These results can be further confirmed when looking at the full 91 species set, where all 25 Platyhelminthes except *Echinostoma caproni* (possibly due to poor genome assembly and annotation) show presence of *dnmt2* orthologs (Figure 2).

Phylogeny of the DNMT2 family fits the species tree (Figure 3), even though nematode DNMT2 has been classified as similar to the vertebrate tRNA methyltransferase DNMT2/TRDMT1 while DNMT2 clearly works as a DNA methyltransferase in Platyhelminthes (Gao et al. 2012). This suggests that the substrate affinity of these proteins may be attributed by small changes in the nucleotide sequence.

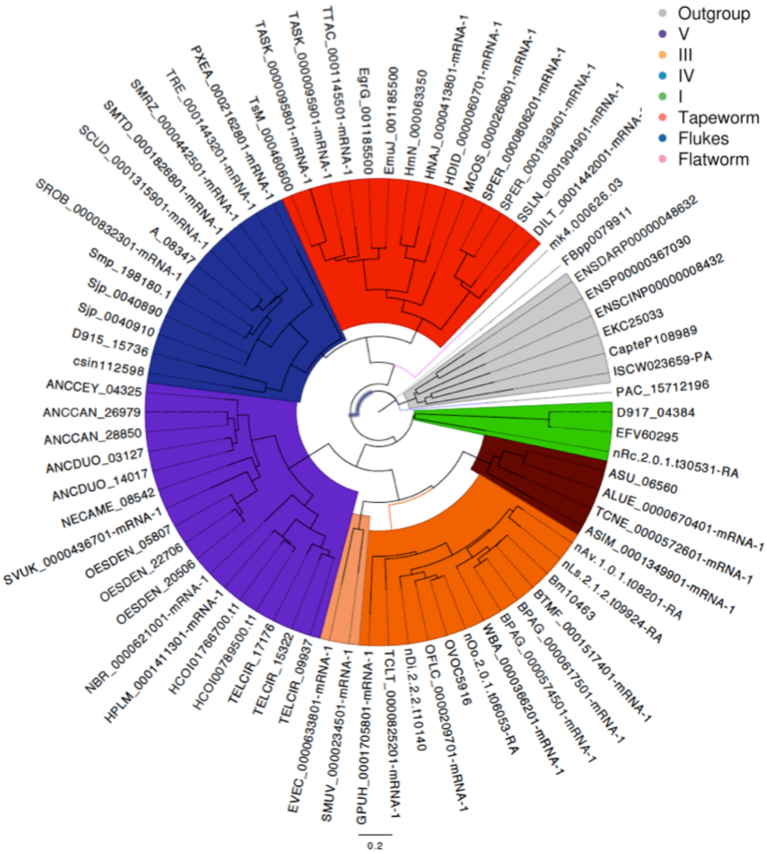


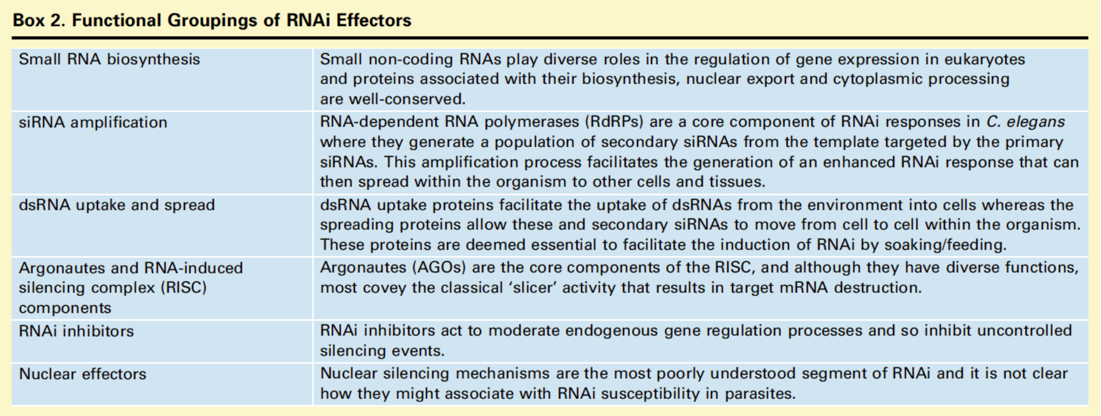
Figure 3. Phylogeny of the DNMT2 family on the 91 50HGP Compara species (based on ClustalW amino acid alignments).

In respect to MBD protein families, there is an increased amount of MBD-related genes in Clade I species (Figure 2), consistent with its potential DNA methylation activity, while in Clade IV MBD proteins are missing, except for a single copy in the plant-parasitic subclade. Platyhelminthes possess a larger repertoire of MBD proteins than even Clade I nematodes.

In summary, this study confirms that cytosine DNA methylation machinery is conserved across Platyhelminthes. Across nematodes only *Trichinella spiralis* (clade I) has been described with functional DNA methylation activity, and here we suggest all clade I nematodes may be capable of cytosine DNA methylation. We also confirm that no other nematodes outside clade I retained DNMT3, the enzyme assumed to be responsible for a functional DNA methylation in this phylum. MBD protein families retained in nematodes and Platyhelminthes are different (except SETMBD2-like family), also suggesting different usage of the DNA methylation machinery between these two phyla.

Presence of RNAi-related genes

Genes known to be involved in *C. elegans* RNAi pathway were taken from (Dalzell et al. 2011; Maule et al. 2011). These include genes involved in different aspects of RNAi (see Box 2 from Dalzell 2011).



Box 2. Functional Groupings of RNAi Effectors

Small RNA biosynthesis	Small non-coding RNAs play diverse roles in the regulation of gene expression in eukaryotes and proteins associated with their biosynthesis, nuclear export and cytoplasmic processing are well-conserved.
siRNA amplification	RNA-dependent RNA polymerases (RdRPs) are a core component of RNAi responses in *C. elegans* where they generate a population of secondary siRNAs from the template targeted by the primary siRNAs. This amplification process facilitates the generation of an enhanced RNAi response that can then spread within the organism to other cells and tissues.
dsRNA uptake and spread	dsRNA uptake proteins facilitate the uptake of dsRNAs from the environment into cells whereas the spreading proteins allow these and secondary siRNAs to move from cell to cell within the organism. These proteins are deemed essential to facilitate the induction of RNAi by soaking/feeding.
Argonautes and RNA-induced silencing complex (RISC) components	Argonautes (AGOs) are the core components of the RISC, and although they have diverse functions, most covey the classical 'slicer' activity that results in target mRNA destruction.
RNAi inhibitors	RNAi inhibitors act to moderate endogenous gene regulation processes and so inhibit uncontrolled silencing events.
Nuclear effectors	Nuclear silencing mechanisms are the most poorly understood segment of RNAi and it is not clear how they might associate with RNAi susceptibility in parasites.

For each of the 73 RNAi-related *C. elegans* genes (73 instead of 77 in the paper because some have been removed or tagged as pseudogenes in wormbase) we identify the Compara gene family where the gene belongs (if any) and analyze the gene presence in our 33 species (the other genes the family are orthologs of the *C. elegans* gene) (Figure 3). This produced 40 Compara families of ortholog genes spanning the 6 functional groupings of RNAi effectors. 8 out of the 73 *C. elegans* genes did not have orthologs with other species.

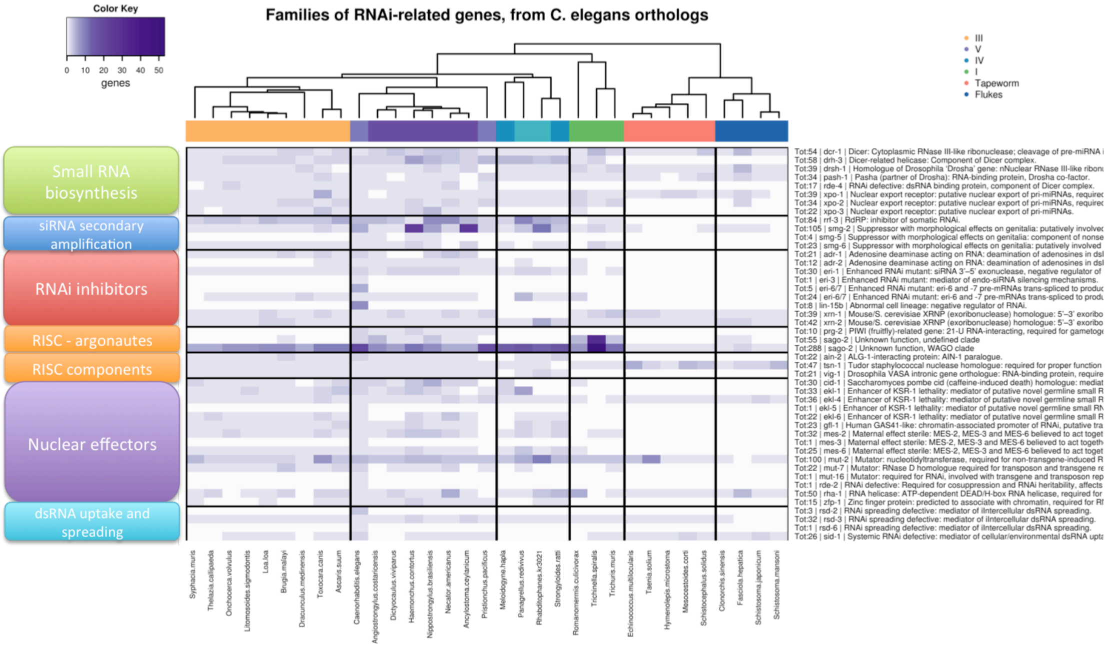


Figure 1. Presence of RNAi-related genes, orthologous to *C. elegans* genes, over the 33 species of interest.

Overall, we observe less conservation of RNAi-related genes in platyhelminths, however this may due to the fact that these are more distant to *C. elegans* than nematodes, as we can also observe better conservation of gene presence in nematode clades V and III, compared to IV and I. Also note that several of the families may perform the same of similar function (e.g. there are several xpo, argonaute and mes families), so the absence of some families in some of the clades may be covered by and increased presence of other families with same function. Having this in mind we observe that all 6 functional categories of RNAi pathway are mostly covered in all species. Notably, the key genes involved in small RNA biosynthesis and the RISC complex (dicer, drosha, pasha, exportin, argonautes) seem to be present (albeit in different copy numbers) in all species under study. However, some functions/genes deemed necessary for effective

RNAi seem to be missing. These missing genes include the *C. elegans* genes rde-4 (dicer-complex), some of the NRDE genes (RISC-complex), vig-1 (RISC component). The lack of piwi argonautes in platyhelminths clustering with the nematode piwi argonautes is described in (Tsai et al. 2013) (and some nematodes; see Sarkies 2015), which is also documented in Zheng 2012 when comparing nematode argonautes platyhelminths. *Trichinella spiralis* shows an expansion of piwi argonautes (check if previously described, perhaps in Bernardo's paper).

To complement our analysis we look for other RNAi-related gene families that might have been missed from using *C. elegans* orthology alone. For this we mined our database of gene families for families with similar pfam domain presence as the families retrieved with the *C. elegans* genes (Figure 4). Note that this was not done extensively and still many RNAi-related families may be missing.

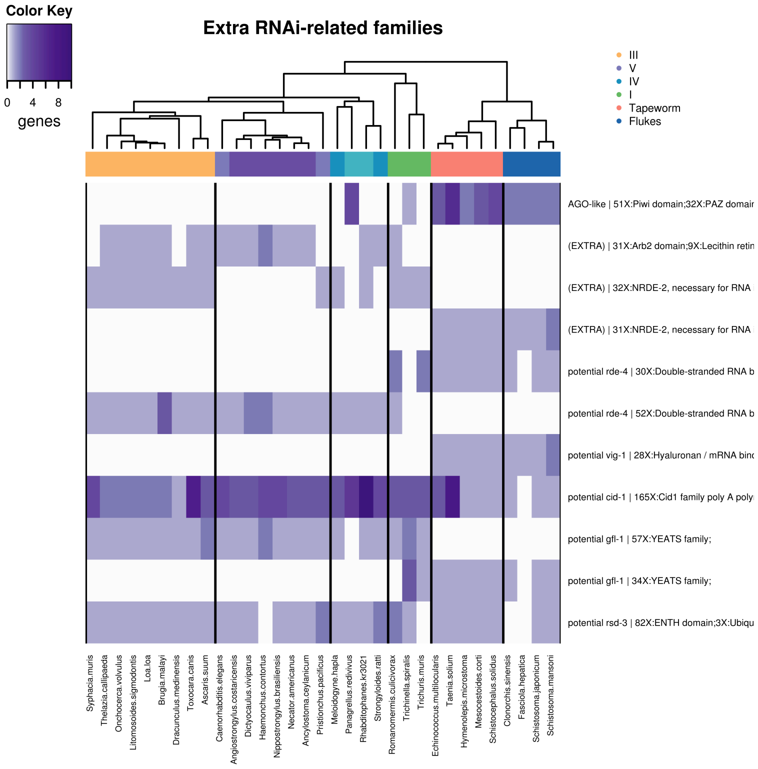


Figure 2. Gene distribution of other potentially RNAi-related families missed from the *C. elegans* orthology-based approach.

Looking for RNAi-related families independently of C. elegans orthlogy allow us to find families of several RNAi effectors that were previously lacking in gene copy numbers, particularly in platyhelminth species. With this approach we find a Piwi-Argonaute-like family enriched in platyhelminthic species, but with a bilaterian root (not shown here, but I have plot with outgroups). This is the family of platyhelminth argonautes described in (Tsai et al. 2013; Zheng 2013) (I actually checked that the genes in Zheng 2013 are the same as in my compare family), but regarded as distant from piwi argonautes of nematodes and planaria. Also, NRDE-like, rde-4-like and vig-1-like families have been found here in platyhelminths.

Data mining for gene expansions and families potentially involved in parasitism

As we cannot closely inspect all the 18,436 families in our dataset (33 high-quality species), we defined ways to highlight the gene families and functions that appear more promising in a parasitism point of view. We collected several metrics to distinguish families (Table 1). Processed by custom perl and python scripts, most of these measures come from the Compara database (all measures related to number and distribution of genes, and gene-tree-based measures such as duplications, losses, roots, branch lengths). All functional annotation-related measures come from an independent InterProScan 5 run for each species.

Previously described gene expansions in parasitic helminthes include genes involved in tissue penetration, digestion of host tissue and evasion of host immune response. Some examples of key parasitism genes include proteases such as chymotrypsin A serine proteases, cysteine, aspartic and metallo-proteases, as well as protease inhibitors (Foth et al. 2014; Laing et al. 2013); genes involved in hemoglobin digestion and anti-coagulant activity (in blood-feeding parasites) (Laing et al. 2013); genes involved in detoxification of potentially damaging compounds such as secreted hydrolases, glutathione peroxidase, fatty acid transporters and ABC transporters (Kikuchi et al. 2011; Tsai et al. 2013; Zarowiecki and Berriman 2014).

Table 1. Gene family metrics for family mining.

familyID -> internal Compara family identifier

Flag -> whether this family should be excluded from analysis (currently from being transposon-related)

#_species -> number of species in each family

#_genes -> number of genes in each family

Mean_genes_per_species -> mean number of genes per species in family

Median_genes_per_species -> median number of genes per species in family

Variation_coefficient_#_genes_per_species -> stdev/mean (of genes per spp)

#_paralogs -> number of paralogs per family = # of genes - # of species

Branch_name -> branch of root of family tree

Completeness_score -> (a.k.a. missingness), % of species (regarding root) present in family

Total_duplications -> total gene duplications in family (compara calculated)

Max_duplications_node -> Species tree node with most duplication events in family

Most_frequent_species -> Species with the highest number of genes in family

caenorhabditis_elegans -> number of genes in this free-living species

panagrellus_redivivus -> number of genes in this free-living species

pristionchus_pacificus -> number of genes in this free-living species

rhabditophanes_kr3021 -> number of genes in this free-living species

schmidtea_mediterranea -> number of genes in this free-living species

Total_losses -> total gene losses in family (compara calculated)

Median_branch_length -> Median branch length of the gene tree (Notes: values >1000 filtered out; this is now for pruned tree

%_genes_with_Pfam -> % of genes in family with at least one pfam domain

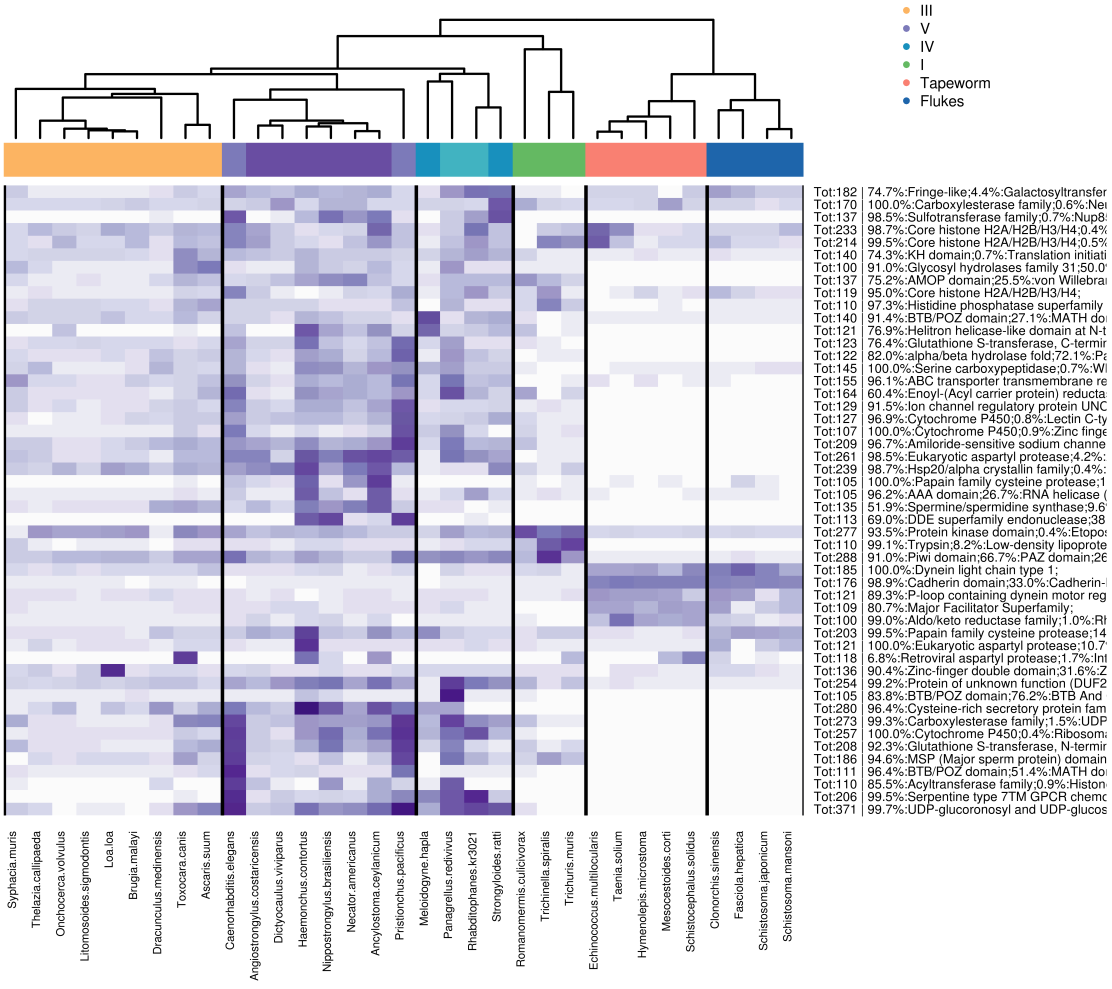
%_genes_with_TMHMM -> % of genes in family with at least one TMHMM hit

%_genes_with_SignalP -> % of genes in family with at least one SinalP_EUK hit

Pfam_perc_in_family -> Percentage of genes in family with at least one of those domains in family (top 3 pfams).

Pfam_all_in_family -> Frequency of each pfam domain found in family, sorted by frequency.

As a first approach to mine our dataset of >18.000 gene families, we searched for clade/species-expanded families by ranking large families (>=100 genes) by their variation coefficient of the number of genes per species, where high values are indicative of gene expansions in a species or a subset of species. We had a closer look at the gene distribution over the 33 species for the 50 large families with higher variation coefficient (Figure 1). Looking at the Pfam domains of those potentially expanded families and performing a GO term enrichment on all these families, we identify several protease families (cysteine, serine, aspartyl), and gene related to carbohydrate/lipid metabolism and response to stress, but also many general-purpose and core biology protein families such as histones, zinc finger, major facilitator superfamily, whose potential role in parasitism is harder to assess.



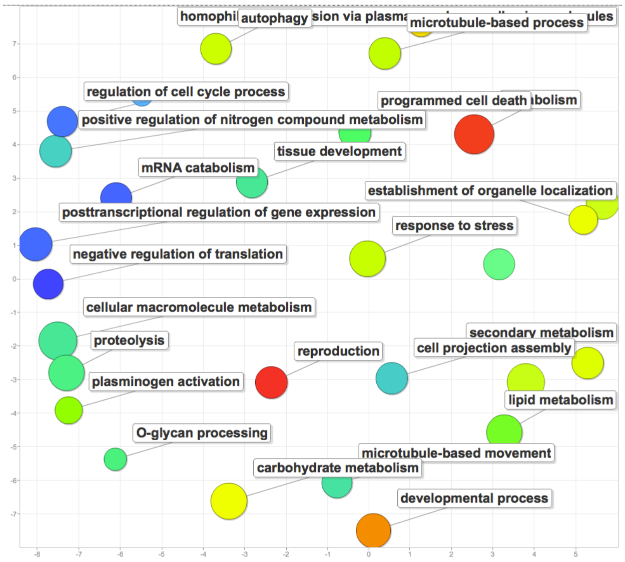


Figure 1. A) Gene distribution of top 50 families with highest variation coefficient and with size more or equal 100 genes. B) Biological process GO term enrichment (TopGO) of all families described in A) (these include genes from all species not just the 33), plotted by REVIGO, reduced to tiny. All p-values <0.05.

Next, as parasites often use excretory/secretory products for host entry, migration and host tissue digestion, we looked for large gene families where >=40% of the genes contain a signal peptide (from SignalPHMM interproscan). This returned 26 families (Figure 2A), several of which are excreted peptidase families, and GO term enrichment analysis return several terms related to response to stimuli, immune response, synaptic transmission and metabolism (Figure 2B). As expected, the cellular components attributed to these genes related to membrane, synaptic and extracellular space (Figure 2C). Many of these families, such as astacins, trypsin, cathepsins, are known to be involved in parasitism or are found to be differentially expressed in parasitic life-stages (Foth et al. 2014; Mak and Ko 2001; Robinson, Dalton, and Donnelly 2008; Williamson et al. 2006). Other families highlighted with this approach are ShK toxin, CAP domain and Transthyretin, which will be explored further.

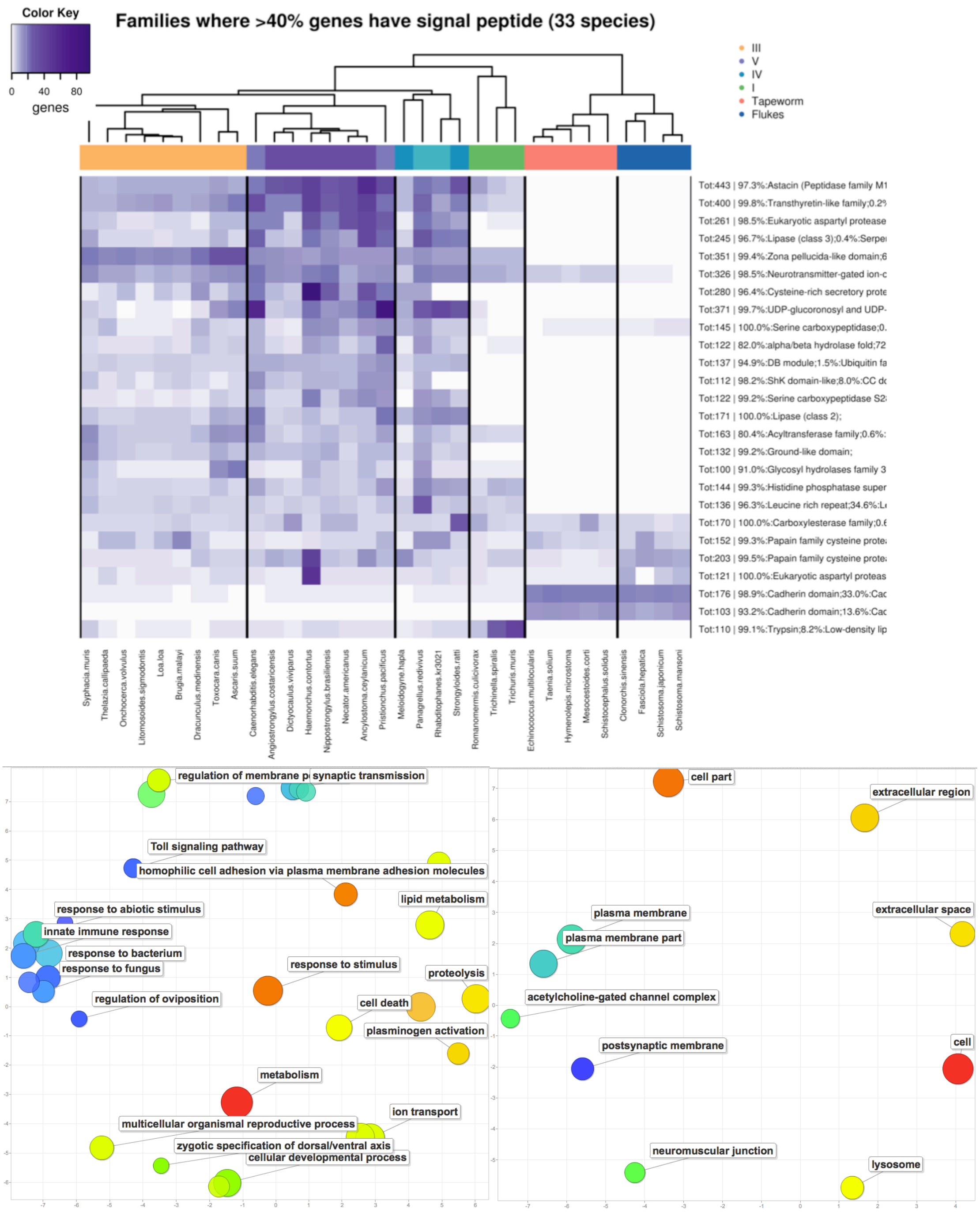
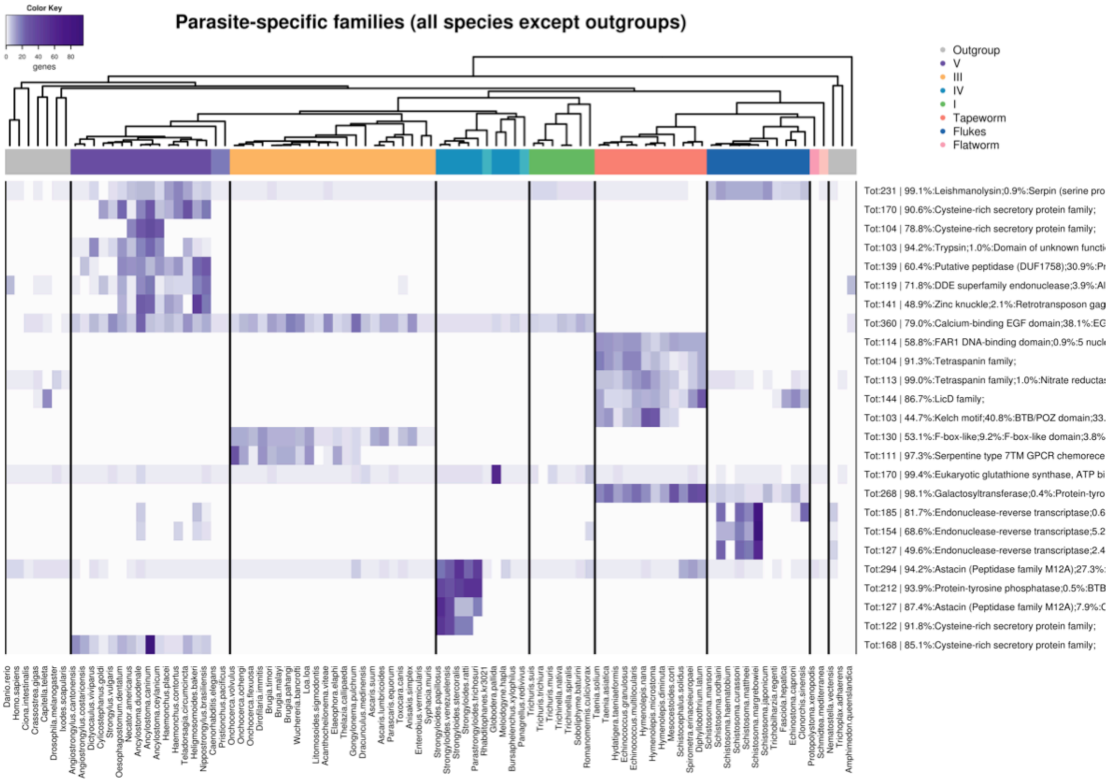
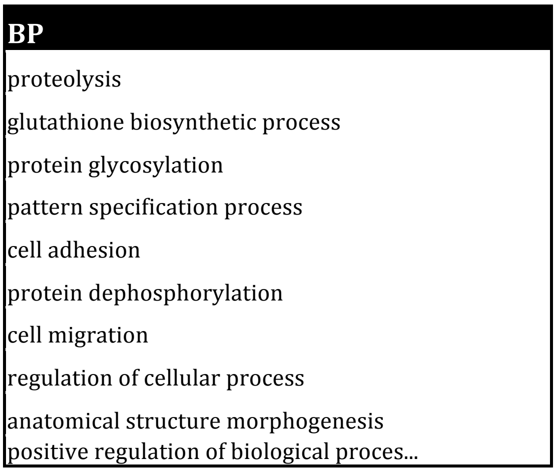


Figure 2. A) Gene distribution of families where >40% contain a signal peptide and with size more or equal 100 genes. B) Biological processes and C) Cell components from GO term enrichment (TopGO) of all families described in A) (these include genes from all species not just the 33), plotted by REVIGO, reduced to large. All p-values <0.05.

Further to the two previous approaches, we looked to find gene families that would be present or expanded on parasitic species and not free-living species. Because our 33 species short-list set do not include any flatworm free-living species, we used the dataset of all species, excluding outgroups (total of 81 species). Families were filtered out by the following criteria: 1) families where any of the 5 free-living species contribute with >0.625% (half the expected in 81-species equality) of the genes in the family, 2) families with less than 100 genes, 3) families where less than 50% of genes have a Pfam annotation. This query returned 25 'parasite-specific' large families (Figure 3A). These families contain many of the previously described proteases involved in parasitism as well as some putative peptidases and highlights several clade specific families (e.g.

FAR1, tetraspanins, F-box-like, LicD/Fukutin). There are also some transposon-related families that have escaped our filters (endonuclease-reverse transcriptase; DDE family pfams are not in our transposon-pfams list).





BP
proteolysis
glutathione biosynthetic process
protein glycosylation
pattern specification process
cell adhesion
protein dephosphorylation
cell migration
regulation of cellular process
anatomical structure morphogenesis
positive regulation of biological proces...

Figure 3. A) Gene distribution of parasite-specific families (described in text). B) Biological process GO term enrichment (TopGO) of all families described in A). All p-values <0.05.

It is worth noting that since we do not have free-living species comparators for every clade, this approach is more prone to find clade-specific families where there is no free-living species in that clade. Also, note that as there are less platyhelminth species than nematode species in our sampling (either full species list or 33 species shortlist), and that we are taking the gene family size into account, we might have increased chance of finding nematode families of interest over platyhelminthic ones.

From these rankings and selections of expanded families, secreted families and parasite-specific families some gene families were highlighted as potentially important for parasites and we sought to explore them further. Note that often species contain several gene families with the same or similar function, either because they originate from different genes, or have diverged enough to not be clustered together, therefore, we will now search for all families with a certain function, using Pfam domain search. These will now be described in separate subtopics.

(note to self: plots were made with different versions of my parser script, meaning some will show number of pfam domains, others will show percentages instead. Also on ealier plots I required at least 6 pfam domains of interest to be present (e.g. >5 astacin domains present in family to be considered 'astacin family'), and later I changed this to at least 10% of genes needing that pfam domain.)
(notes/methods about the plots (all of them except RNAi and DNA methylation: plots use hclust with euclidian distance function for the family sorting/clustering, so that we can better visualize of clade-specific families. The lighter colours on the species/clades legend represent free-living species. All gene numbers are plain copy numbers (not transformed in any way), however the color progression is scaled in a way that the colour looks similar when high numbers, but contrast is higher with low numbers. Plots from by R gplots library, heatmap.2 function, wrapped in python code.)

Astacin metallopeptidases

Astacins are a family of metallopeptidases belonging to the MEROPS peptidase family M12. They are usually secreted or associated with plasma membrane and are involved in activation of growth factors, processing of extracellular proteins and degradation of polypeptides (Dumermuth et al. 1991). Astacin expression have been identified in infective life-states of nematodes from different clades and are regarded as important for nematode parasitism (Borchert et al. 2007; Mak and Ko 2001; Williamson et al. 2006)) (V. Hunt Strongyloides manuscript in preparation).

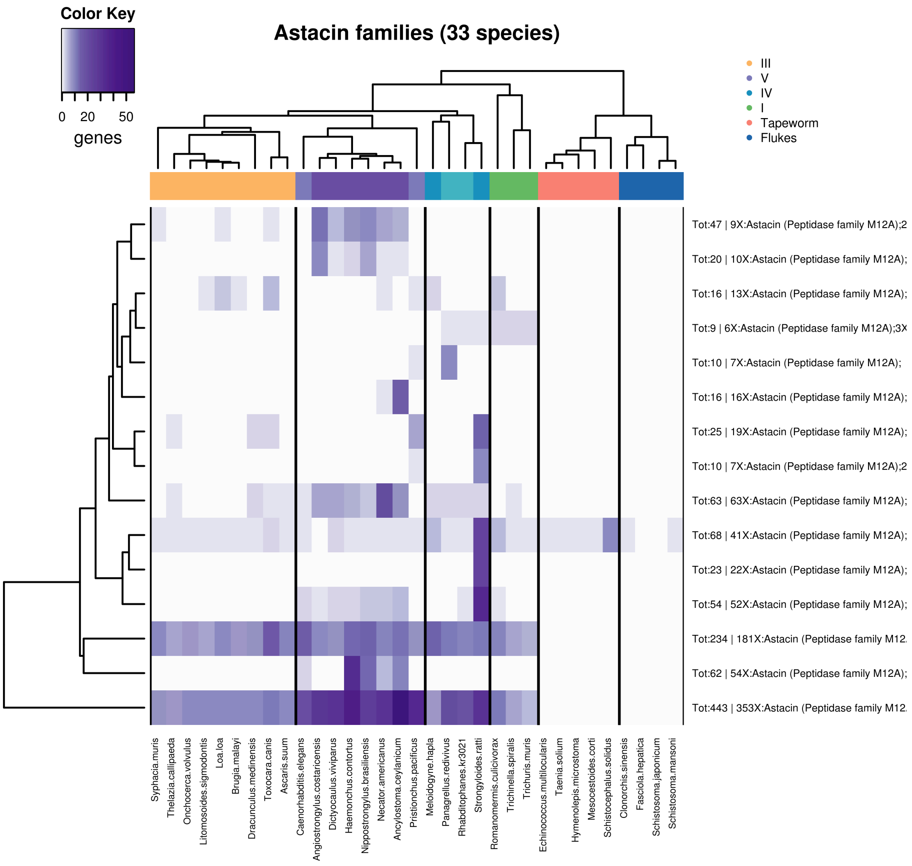


Figure 4. Gene distribution of candidate Astacin families. These were found by searching for families with 'Astacin' Pfam domain in the '33_shortlisted_species' table. Families with <6 Astacin domains were filtered out.

A total of 15 potential astacin families, comprising a total of more than 1000 genes were found in the 33 high-quality genomes. 443 of these genes are clustered in a single family enriched in clade V species, but several other families, specific to clade IV are found (Figure 4). Platyhelminthes are not enriched for these peptidases.

Trypsin serine proteases

Trypsin (EC 3.4.21.4) is a serine protease, hydrolysing proteins. These have been identified in Apicomplexan parasites (Arenas et al. 2010), and also expression of chymotrypsin A-like serine proteases has been described in *Trichuris muris* nematode (Foth et al. 2014). Also, trypsin inhibitors have been suggested as an important factor in Ascaris species specificity (Hawley and Peanasky 1992).

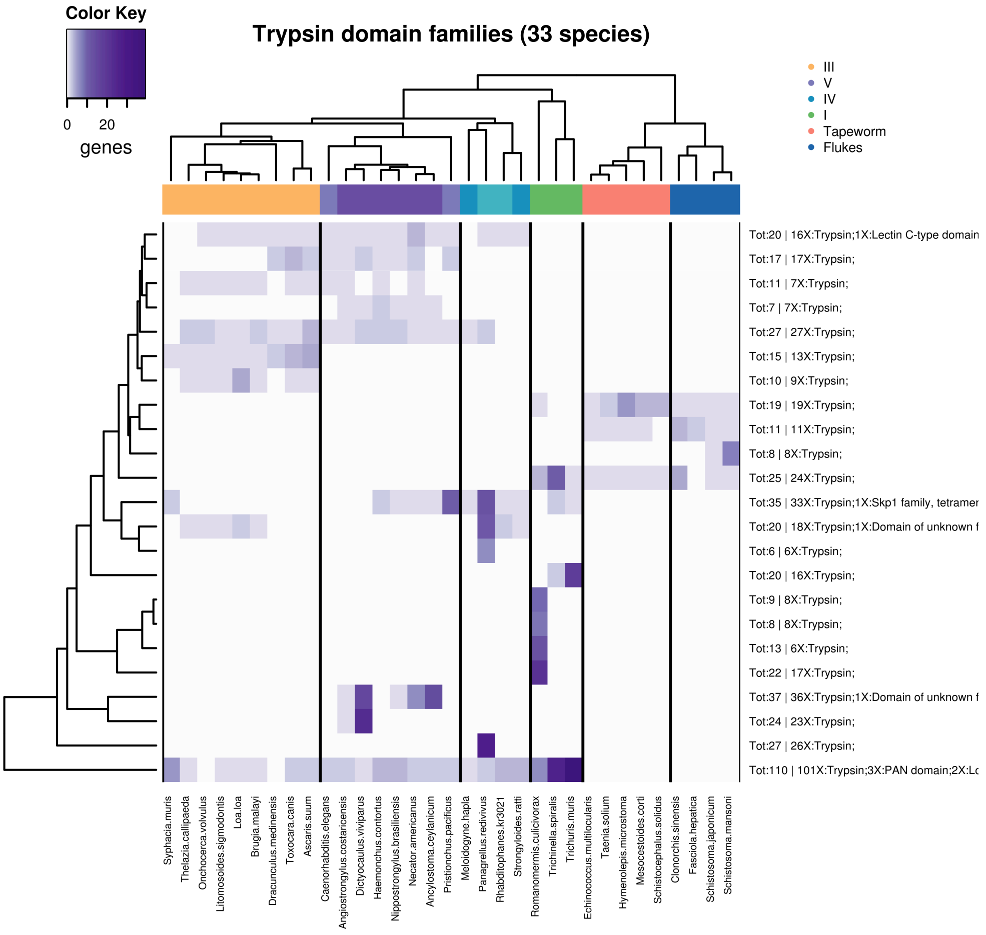


Figure 5. Gene distribution of candidate Trypsin families. These were found by searching for families with 'Trypsin' Pfam domain description in the '33_shortlisted_species' table. Families with <6 Trypsin domains were filtered out.

This search harboured a total of 501 candidate Trypsin genes clustered in 23 families (Figure 5). These seem particularly expanded in Clade I species (described in *T. muris* (Foth et al. 2014)), particularly *Romanomermis culcivorax*, but also present in large numbers in *Dyctioculus viviparus* and *Panagrellus redivivus* (a free-living nematode).

Serpin (serine protease inhibitor)

Serpin name comes from its usual function as inhibitor of chymotrypsin-like serine proteases. This proteins have been described in helminth infections, used to protect against the host proteases, in *Trichostrongylus vitrines* (a strongylid nematode, black scour worm) and *Schistosoma japonicum* (Knox 2007; MacLennan, McLean, and Knox 2005; Yan et al. 2005).

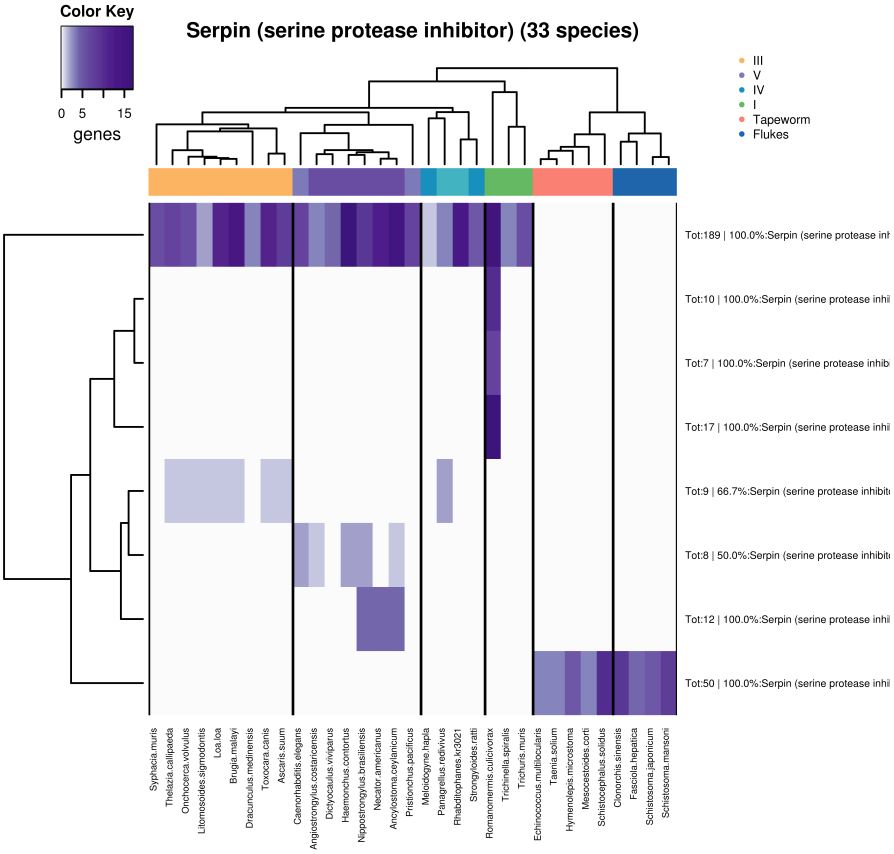


Figure 6. Gene distribution of candidate Serpin families. These were found by searching for families with 'Serpin' Pfam domain description in the '33_shortlisted_species' table. Families where less than 10% of genes had Serpin domains were filtered out.

We find a total of 302 Serpin genes clustered into 8 families (Figure 6). Two large families, one of nematodes and another of Platyhelminthes, these genes have been described in both phyla. *Romanomermis culcivorax* presents three other species-specific Serpin families.

Papain family cysteine proteases (cathepsin)

Cysteine proteases (thiol proteases) are commonly found in fruits, including papaya, pineapple and figs. These proteins have been associated to host entry, tissue migration and suppression of immune responses in several platyhelminth parasites (Robinson, Dalton, and Donnelly 2008).

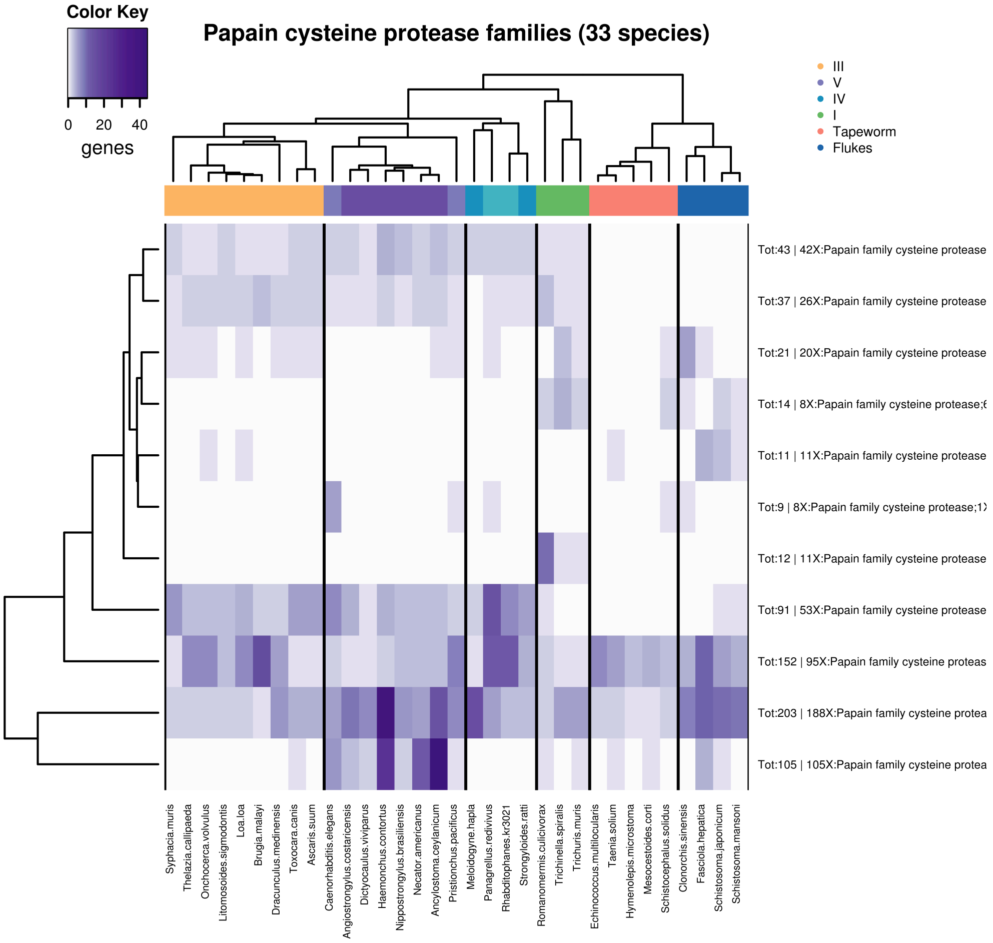


Figure 7. Gene distribution of candidate Papain cysteine protease families. These were found by searching for families with ‘Papain’ Pfam domain in the '33_shortlisted_species' table. Families with <6 Papain domains were filtered out.

This search returned 698 genes in 11 families (Figure 7). The *Fasciola hepatica* and trematode families previously described were found here, but more interestingly, papain cysteine proteases are highly prevalent in nematode species as well, especially in *Haemonchus contortus* and *Ancylostoma ceylanicum*.

Eukaryotic aspartyl protease

Aspartic proteases use aspartate residues for the catalysis of peptides, having optimal activity in acid pH. Families of these proteases have been found in several Strongyloides species, *Onchocerca volvulus*, *Brugia malayi* and strongylid blood feeders such as *Haemonchus contortus*, *Ancylostoma caninum* and *Necator americanus*, where its role in haemoglobin digestion has been described (Mello et al. 2009). In hookworms, aspartic proteases have also been associated to skin penetration by degrading skin macromolecules (McKerrow et al. 1990).

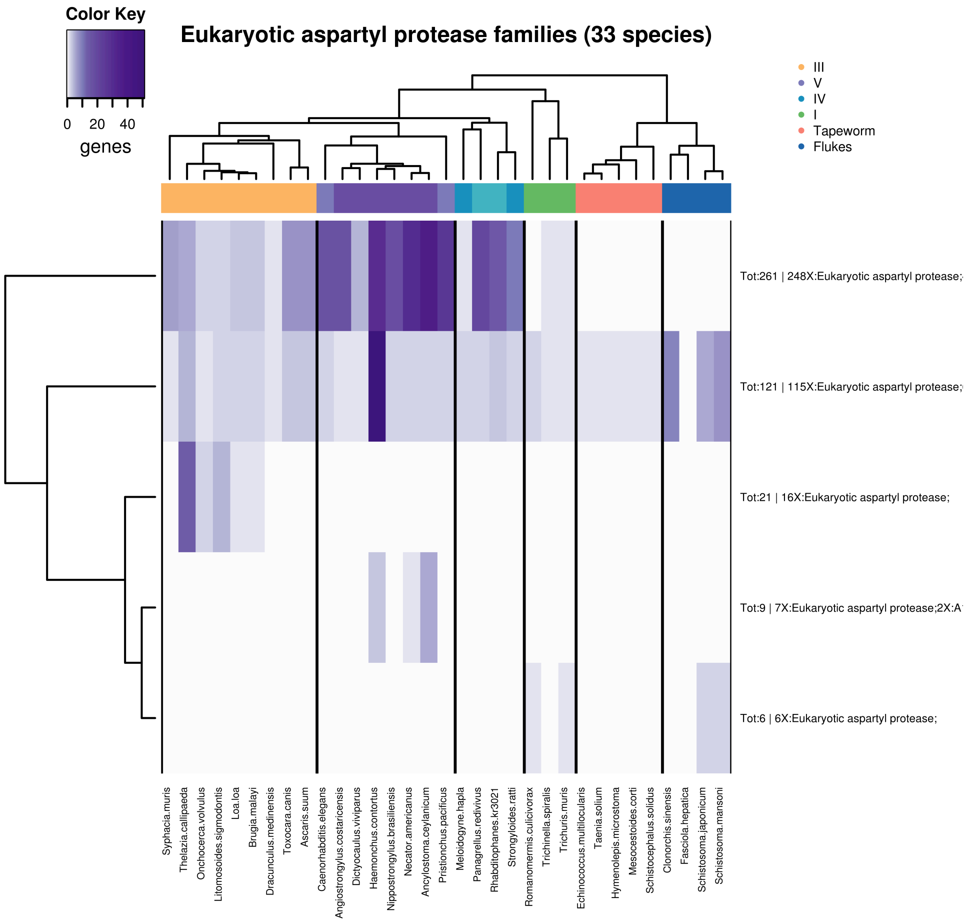


Figure 8. Gene distribution of candidate Aspartyl protease families. These were found by searching for families with ‘Eukaryotic Aspartyl protease’ Pfam domain in the '33_shortlisted_species' table. Families with <6 Aspartyl domains were filtered out.

We find 418 potential aspartyl proteases spread over 5 families (Figure 8). Blood feeding species (*H. contortus, A. ceylanicum, N. brasiliensis, N. necator*) are especially enriched in these proteases (except *D. viviparus*, but is this really a blood-feeder?). Also clade IV species contain several of these proteases.

Cysteine-rich secretory protein family (CAP-domain, SCP/TAPS)

CAP protein families (**c**ysteine-rich secretory proteins, **a**ntigen 5, and **p**athogenesis-related 1 proteins) are found both in prokaryotes and eukaryotes. Subfamilies of these proteins include Golgi-associated pathogenesis related (GAPR1), peptidase inhibitors 15 and 16, cysteine-rich secretory proteins (CRISPs). These proteins are often secreted, and show to be expressed in strongylid nematodes (and Strongyloides) during host invasion (Cantacessi and Gasser 2012).

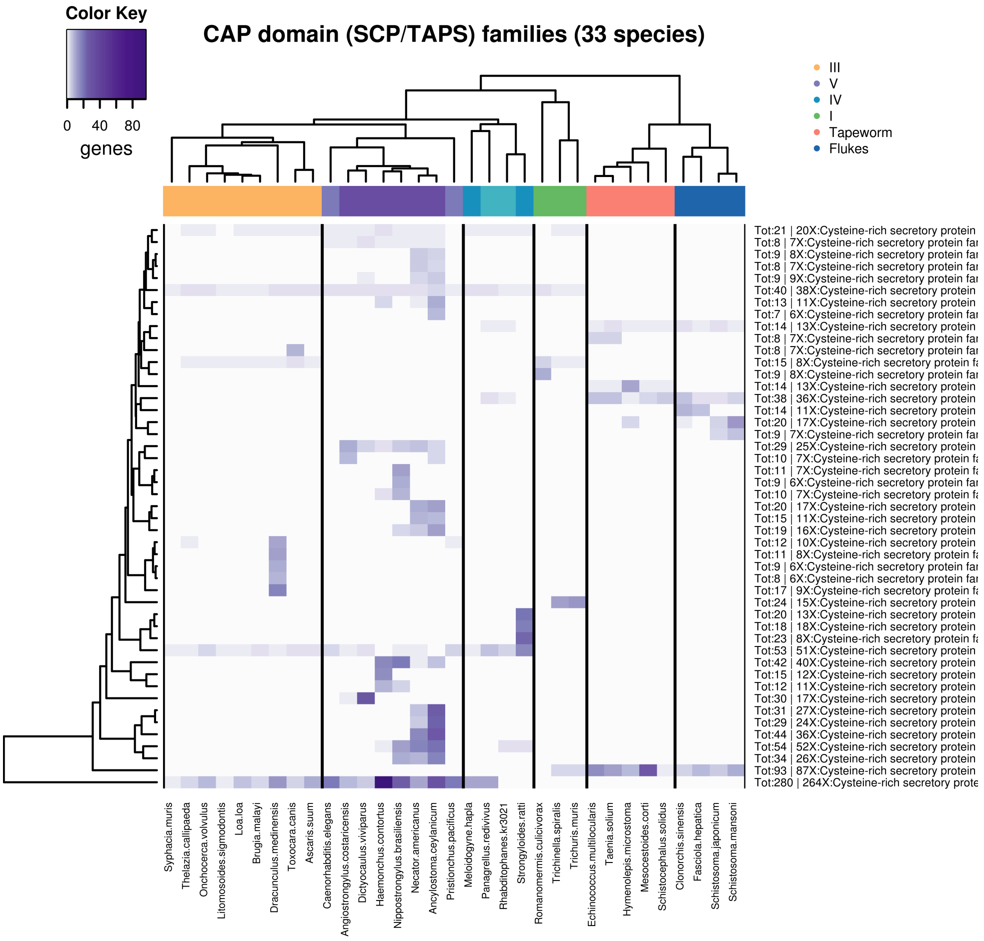


Figure 9. Gene distribution of candidate CAP domain (SCP/TAPS) families. These were found by searching for families with 'Cysteine-rich secretory protein family' Pfam domain in the '33_shortlisted_species' table. Families with <6 CAP domains were filtered out.

CAP domain families are spread over 47 families (Figure 9), suggesting sequence similarity of these genes is low, and gene expansions occurring independently, even within the same species. These families are particularly enriched in parasitic clade V and parasitic clade IV nematodes, as previously described.

Transthyretin-like families

HIUase/Transthyretin family original function is HIU hydrolysis. Transthyretin-like is a nematode-specific family that has weak similarity to transthyretin, and its function is unknown. However, these have been described in plant-parasite *Radopholus similis* and other nematode parasites (being expressed at parasitic stages) (Jacob et al. 2007; Saverwyns et al. 2008), and also found in *C. elegans*, where it mediates recognition of apoptotic cells (Wang et al. 2010).

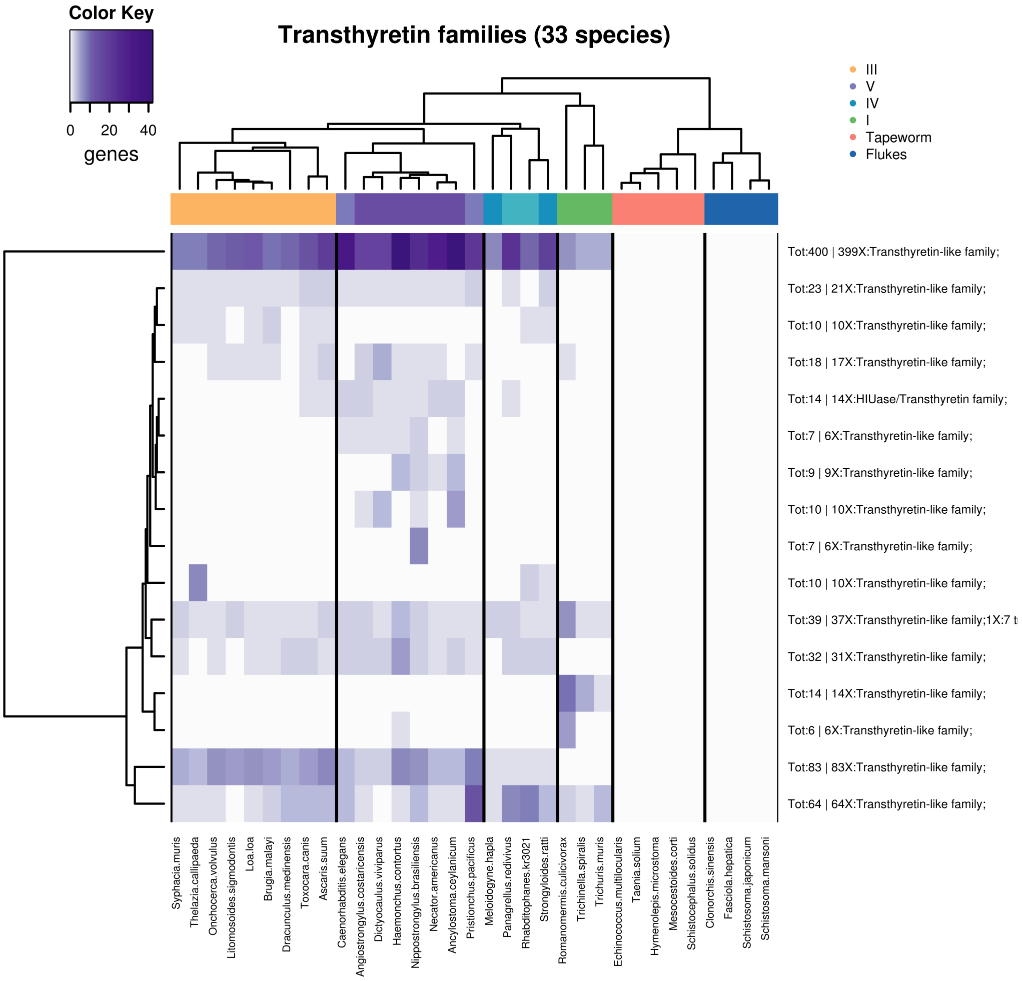


Figure 10. Gene distribution of candidate Transthyretin families. These were found by searching for families with 'Transthyretin' Pfam domain in the '33_shortlisted_species' table. Families with <6 Transthyretin domains were filtered out. Note that this not only includes Transthyretin-like families but also contains a HIUase/Transthyretin family with 14 members.

15 families of transthyretin-like (732 genes) and 1 family of HIUase/Transthyretin (14 genes) are found (Figure 10). We confirm that transthyretin-like families are exclusive to nematode species, and show to be prevalent in all nematodes, more expanded in clade V, but also in free-living species.

DNase II (deoxyribonuclease II)

DNases degrade DNA by hydrolysing phosphodiester linkages in the DNA backbone. DNases are present in metazoans and DNase II has been found highly enriched in the *Trichinella spiralis* nematode (166 copies), expressed at high levels in the anterior region, which is in contact with the cytoplasm of the host cells (Foth et al. 2014).

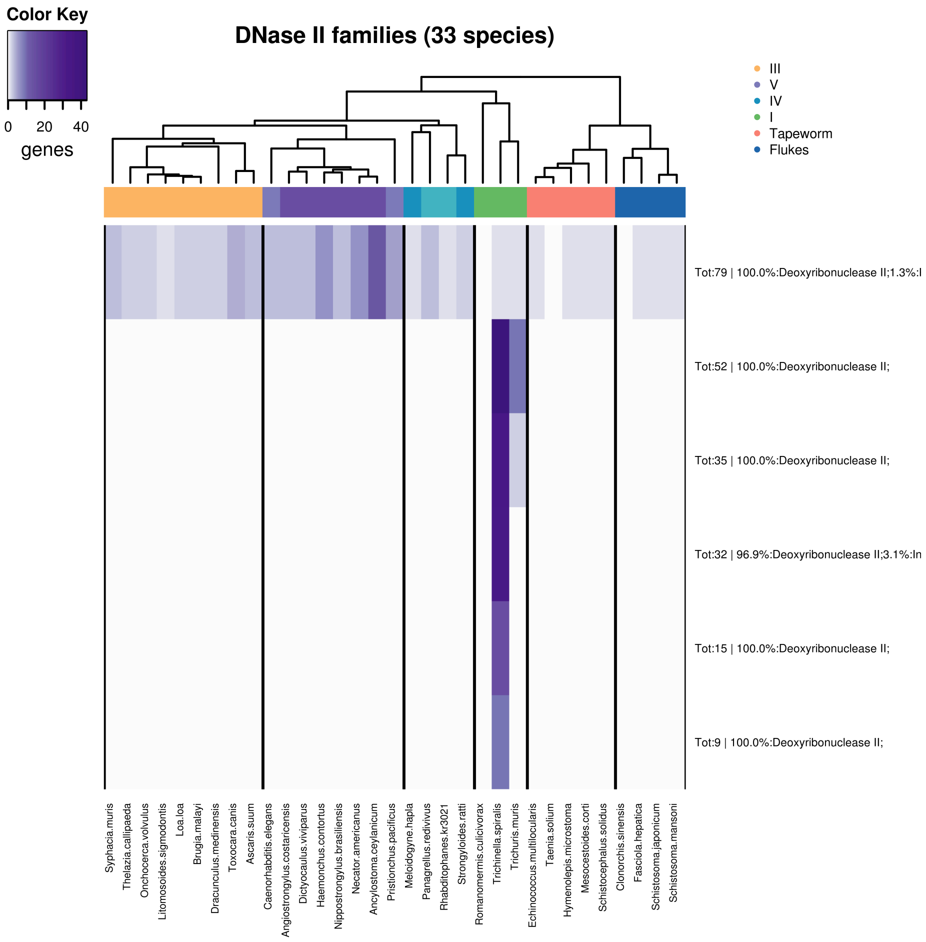


Figure 11. Gene distribution of candidate DNase II families. These were found by searching for families with 'Deoxyribonuclease II' Pfam domain in the '33_shortlisted_species' table. Families where less than 10% of genes had DNase II domains were filtered out.

We find 140 DNase II genes in *T. spiralis*, short of the 166 described in (Foth et al. 2014), and conclude that this expansion is exclusive for Trichuris species.

Stichodactyla (ShK) toxin

Stichodactyla toxin is a peptide toxin that blocks voltage-gated potassium channels (Kv). It has been suggested that helminth ShK-related peptides are associated with the beneficial effects of probiotic helminth therapy in human autoimmune diseases (Chhabra et al. 2014).

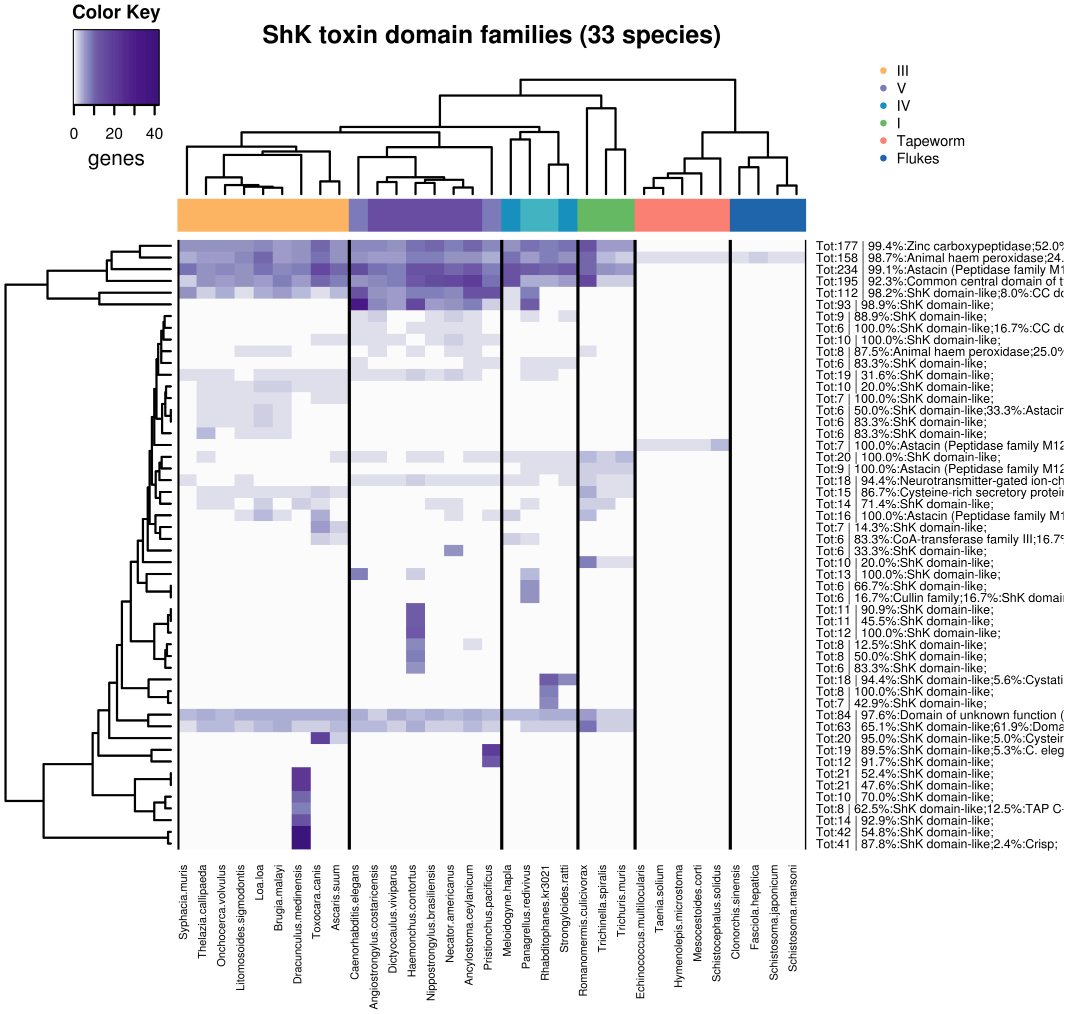


Figure 12. Gene distribution of candidate ShK families. These were found by searching for families with 'ShK domain' Pfam domain in the '33_shortlisted_species' table. Families where less than 10% of genes had ShK domains were filtered out.

We found a total of 1659 genes in 54 families, most present in nematode species (Figure 12). ShK genes presence in several clades of nematodes and in *S. mansoni* has been described. Note that the first 4 families, with a nematode-wide presence, contain more than 10% genes with ShK domain but also contain other domains, likely conferring other functions. Focusing on the remaining families we observe an enrichment of ShK genes in *Dracunculus medinensis* and *Haemonchus contortus*.

Antequera, Francisco, Mercedes Tamame, and Julio R Villanuevaz. 1984. "DNA Methylation in the Fungi*." : 8033–36.

Borchert, Nadine, Christoph Becker-Pauly, Antje Wagner, Peter Fischer, Walter Stöcker, and Norbert W. Brattig. 2007. "Identification and Characterization of Onchoastacin, an Astacin-like Metalloproteinase from the Filaria Onchocerca Volvulus." *Microbes and Infection* 9: 498–506.

Cantacessi, Cinzia, and Robin B. Gasser. 2012. "Scp/taps Proteins in Helminths - Where to from Now?" *Molecular and Cellular Probes* 26: 54–59.

Capuano, Floriana, Robert Kok, Henk J Blom, and Markus Ralser. 2014. “Cytosine DNA Methylation Is Found in Drosophila Melanogaster but Absent in Saccharomyces Cerevisiae, Schizosaccharomyces Pombe , and Other Yeast Species.”

Chhabra, Sandeep, Shih Chieh Chang, Hai M Nguyen, Redwan Huq, Mark R Tanner, Luz M Londono, Rosendo Estrada, Vikas Dhawan, Satendra Chauhan, Sanjeev K Upadhyay, Mariel Gindin, Peter J Hotez, Jesus G Valenzuela, Biswaranjan Mohanty, James D Swarbrick, Heike Wulff, Shawn P Iadonato, George a Gutman, Christine Beeton, Michael W Pennington, Raymond S Norton, and K George Chandy. 2014. “Kv1.3 Channel-Blocking Immunomodulatory Peptides from Parasitic Worms: Implications for Autoimmune Diseases.” *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 28(9): 3952–64. http://www.ncbi.nlm.nih.gov/pubmed/24891519 (June 5, 2015).

Dalzell, Johnathan J, Paul McVeigh, Neil D Warnock, Makedonka Mitreva, David McK Bird, Pierre Abad, Colin C Fleming, Tim a Day, Angela Mousley, Nikki J Marks, and Aaron G Maule. 2011. “RNAi Effector Diversity in Nematodes.” *PLoS neglected tropical diseases* 5(6): e1176. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3110158&tool=pmcentrez&rendertype=abstract (June 3, 2015).

Dumermuth, E., E. E. Sterchi, W. Jiang, R. L. Wolz, J. S. Bond, A. V. Flannery, and R. J. Beynon. 1991. “The Astacin Family of Metalloendopeptidases.” *Journal of Biological Chemistry* 266: 21381–85.

Flicek, Paul, Ikhlak Ahmed, M Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, Peter Clapham, Guy Coates, Susan Fairley, Stephen Fitzgerald, Laurent Gil, Carlos García-Girón, Leo Gordon, Thibaut Hourlier, Sarah Hunt, Thomas Juettemann, Andreas K Kähäri, Stephen Keenan, Monika Komorowska, Eugene Kulesha, Ian Longden, Thomas Maurel, William M McLaren, Matthieu Muffato, Rishi Nag, Bert Overduin, Miguel Pignatelli, Bethan Pritchard, Emily Pritchard, Harpreet Singh Riat, Graham R S Ritchie, Magali Ruffier, Michael Schuster, Daniel Sheppard, Daniel Sobral, Kieron Taylor, Anja Thormann, Stephen Trevanion, Simon White, Steven P Wilder, Bronwen L Aken, Ewan Birney, Fiona Cunningham, Ian Dunham, Jennifer Harrow, Javier Herrero, Tim J P Hubbard, Nathan Johnson, Rhoda Kinsella, Anne Parker, Giulietta Spudich, Andy Yates, Amonida Zadissa, and Stephen M J Searle. 2013. “Ensembl 2013.” *Nucleic acids research* 41(Database issue): D48–55. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531136&tool=pmcentrez&rendertype=abstract (December 1, 2014).

Foth, Bernardo J, Isheng J Tsai, Adam J Reid, Allison J Bancroft, Sarah Nichol, Alan Tracey, Nancy Holroyd, James a Cotton, Eleanor J Stanley, Magdalena Zarowiecki, Jimmy Z Liu, Thomas Huckvale, Philip J Cooper, Richard K Grencis, and Matthew Berriman. 2014. “Whipworm Genome and Dual-Species Transcriptome Analyses Provide Molecular Insights into an Intimate Host-Parasite Interaction.” *Nature genetics* 46(7): 693–700. http://www.ncbi.nlm.nih.gov/pubmed/24929830 (June 3, 2015).

Gao, Fei, Xiaolei Liu, Xiu-Ping Wu, Xue-Lin Wang, Desheng Gong, Hanlin Lu, Yudong Xia, Yanxia Song, Junwen Wang, Jing Du, Siyang Liu, Xu Han, Yizhi Tang, Huanming Yang, Qi Jin, Xiuqing Zhang, and Mingyuan Liu. 2012. “Differential DNA Methylation in Discrete Developmental Stages of the Parasitic Nematode Trichinella Spiralis.”

Genome biology 13(10): R100. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053732&tool=pmcentrez&rendertype=abstract (June 3, 2015).

Gao, Fei, Rui Wang, and Mingyuan Liu. 2014. "Trichinella Spiralis, Potential Model Nematode for Epigenetics and Its Implication in Metazoan Parasitism." *Frontiers in physiology* 4(January): 410. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3887316&tool=pmcentrez&rendertype=abstract (June 3, 2015).

Geyer, Kathrin K, Iain W Chalmers, Neil Mackintosh, Julie E Hirst, Rory Geoghegan, Mathieu Badets, Peter M Brophy, Klaus Brehm, and Karl F Hoffmann. 2013. "Cytosine Methylation Is a Conserved Epigenetic Feature Found throughout the Phylum Platyhelminthes." *BMC genomics* 14(1): 462. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3710501&tool=pmcentrez&rendertype=abstract (June 3, 2015).

Goll, Mary Grace, Finn Kirpekar, Keith A Maggert, Jeffrey A Yoder, Chih-lin Hsieh, Xiaoyu Zhang, Kent G Golic, Steven E Jacobsen, and Timothy H Bestor. 2006. "Methylation of tRNA Asp by the DNA Methyltransferase Homolog Dnmt2." 311(January): 395–98.

Hu, Chiung-Wen, Jian-Lian Chen, Yu-Wen Hsu, Cheng-Chieh Yen, and Mu-Rong Chao. 2015. "Trace Analysis of Methylated and Hydroxymethylated Cytosines in DNA by Isotope-Dilution LC-MS/MS: First Evidence of DNA Methylation in Caenorhabditis Elegans." *The Biochemical journal* 465(1): 39–47. http://www.ncbi.nlm.nih.gov/pubmed/25299492 (May 21, 2015).

Jacob, Joachim, Bartel Vanholme, Annelies Haegeman, and Godelieve Gheysen. 2007. "Four Transthyretin-like Genes of the Migratory Plant-Parasitic Nematode Radopholus Similis: Members of an Extensive Nematode-Specific Family." *Gene* 402: 9–19.

Kikuchi, Taisei, James a Cotton, Jonathan J Dalzell, Koichi Hasegawa, Natsumi Kanzaki, Paul McVeigh, Takuma Takanashi, Isheng J Tsai, Samuel a Assefa, Peter J a Cock, Thomas Dan Otto, Martin Hunt, Adam J Reid, Alejandro Sanchez-Flores, Kazuko Tsuchihara, Toshiro Yokoi, Mattias C Larsson, Johji Miwa, Aaron G Maule, Norio Sahashi, John T Jones, and Matthew Berriman. 2011. "Genomic Insights into the Origin of Parasitism in the Emerging Plant Pathogen Bursaphelenchus Xylophilus." *PLoS pathogens* 7(9): e1002219. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3164644&tool=pmcentrez&rendertype=abstract (June 3, 2015).

Knox, D. P. 2007. "Proteinase Inhibitors and Helminth Parasite Infection." *Parasite Immunology* 29: 57–71.

Laing, Roz, Taisei Kikuchi, Axel Martinelli, Isheng J Tsai, Robin N Beech, Elizabeth Redman, Nancy Holroyd, David J Bartley, Helen Beasley, Collette Britton, David Curran, Eileen Devaney, Aude Gilabert, Martin Hunt, Frank Jackson, Stephanie L Johnston, Ivan Kryukov, Keyu Li, Alison a Morrison, Adam J Reid, Neil Sargison, Gary I Saunders, James D Wasmuth, Adrian Wolstenholme, Matthew Berriman, John S Gilleard, and James a Cotton. 2013. "The Genome and Transcriptome of Haemonchus Contortus, a Key Model Parasite for Drug and Vaccine Discovery."

Genome biology 14(8): R88. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4054779&tool=pmcentrez&rendertype=abstract (June 3, 2015).

Law, Julie A, and Steven E Jacobsen. 2011. “Patterns in Plants and Animals.” 11(3): 204–20.

MacLennan, K, K McLean, and D P Knox. 2005. “Serpin Expression in the Parasitic Stages of Trichostrongylus Vitrinus, an Ovine Intestinal Nematode.” *Parasitology* 130: 349–57.

Mak, Chi-ho, and Ronald C. Ko. 2001. “Characterization of Endonuclease Activity from Excretory/secretory Products of a Parasitic Nematode, Trichinella Spiralis.” *European Journal of Biochemistry* 260(2): 477–81. http://doi.wiley.com/10.1046/j.1432-1327.1999.00174.x.

Manuscript, Author. 2013. “Epigenetic Alterations in Oncogenesis” ed. Adam R. Karpf. 754: 1–25. http://link.springer.com/10.1007/978-1-4419-9967-2 (March 4, 2015).

Maule, Aaron G, Paul McVeigh, Johnathan J Dalzell, Louise Atkinson, Angela Mousley, and Nikki J Marks. 2011. “An Eye on RNAi in Nematode Parasites.” *Trends in parasitology* 27(11): 505–13. http://www.ncbi.nlm.nih.gov/pubmed/21885343 (June 3, 2015).

McKerrow, J H, P Brindley, M Brown, A A Gam, C Staunton, and F A Neva. 1990. “Strongyloides Stercoralis: Identification of a Protease That Facilitates Penetration of Skin by the Infective Larvae.” *Experimental parasitology* 70: 134–43.

Mello, Luciane V, Helen O’Meara, Daniel J Rigden, and Steve Paterson. 2009. “Identification of Novel Aspartic Proteases from Strongyloides Ratti and Characterisation of Their Evolutionary Relationships, Stage-Specific Expression and Molecular Structure.” *BMC genomics* 10: 611.

Raddatz, Günter, Paloma M Guzzardo, Nelly Olova, Marcelo Rosado Fantappié, Markus Rampp, Matthias Schaefer, Wolf Reik, Gregory J Hannon, and Frank Lyko. 2013. “Dnmt2-Dependent Methylomes Lack Defined DNA Methylation Patterns.” *Proceedings of the National Academy of Sciences of the United States of America* 110(21): 8627–31. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3666705&tool=pmcentrez&rendertype=abstract (May 16, 2015).

Robinson, Mark W., John P. Dalton, and Sheila Donnelly. 2008. “Helminth Pathogen Cathepsin Proteases: It’s a Family Affair.” *Trends in Biochemical Sciences* 33: 601–8.

Roloff, Tim C, H Hilger Ropers, and Ulrike A Nuber. 2003. “Comparative Study of Methyl-CpG-Binding Domain Proteins.” 9: 1–9.

Saverwyns, H., A. Visser, J. Van Durme, D. Power, I. Morgado, M. W. Kennedy, D. P. Knox, J. Schymkowitz, F. Rousseau, K. Gevaert, J. Vercruysse, E. Claerebout, and P. Geldhof. 2008. “Analysis of the Transthyretin-like (TTL) Gene Family in Ostertagia Ostertagi - Comparison with Other Strongylid Nematodes and Caenorhabditis Elegans.” *International Journal for Parasitology* 38: 1545–56.

Schaefer, Matthias, and Frank Lyko. 2010. “Solving the Dnmt2 Enigma.” *Chromosoma* 119(1): 35–40. http://www.ncbi.nlm.nih.gov/pubmed/19730874 (June 3, 2015).

Schiffer, Philipp H, Michael Kroiher, Christopher Kraus, Georgios D Koutsovoulos, Sujai Kumar, Julia I R Camps, Ndifon a Nsah, Dominik Stappert, Krystalynne Morris, Peter Heger, Janine Altmüller, Peter Frommolt, Peter Nürnberg, W Kelley Thomas, Mark L Blaxter, and Einhard Schierenberg. 2013. “The Genome of Romanomermis Culicivorax: Revealing Fundamental Changes in the Core Developmental Genetic Toolkit in Nematoda.” *BMC genomics* 14: 923. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3890508&tool=pmcentrez&rendertype=abstract.

Tsai, Isheng J, Magdalena Zarowiecki, Nancy Holroyd, Alejandro Garciarrubio, Alejandro Sanchez-Flores, Karen L Brooks, Alan Tracey, Raúl J Bobes, Gladis Fragoso, Edda Sciutto, Martin Aslett, Helen Beasley, Hayley M Bennett, Jianping Cai, Federico Camicia, Richard Clark, Marcela Cucher, Nishadi De Silva, Tim a Day, Peter Deplazes, Karel Estrada, Cecilia Fernández, Peter W H Holland, Junling Hou, Songnian Hu, Thomas Huckvale, Stacy S Hung, Laura Kamenetzky, Jacqueline a Keane, Ferenc Kiss, Uriel Koziol, Olivia Lambert, Kan Liu, Xuenong Luo, Yingfeng Luo, Natalia Macchiaroli, Sarah Nichol, Jordi Paps, John Parkinson, Natasha Pouchkina-Stantcheva, Nick Riddiford, Mara Rosenzvit, Gustavo Salinas, James D Wasmuth, Mostafa Zamanian, Yadong Zheng, Xuepeng Cai, Xavier Soberón, Peter D Olson, Juan P Laclette, Klaus Brehm, and Matthew Berriman. 2013. “The Genomes of Four Tapeworm Species Reveal Adaptations to Parasitism.” *Nature* 496(7443): 57–63. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3964345&tool=pmcentrez&rendertype=abstract (June 3, 2015).

Wang, Xiaochen, Weida Li, Dongfeng Zhao, Bin Liu, Yong Shi, Baohui Chen, Hengwen Yang, Pengfei Guo, Xin Geng, Zhihong Shang, Erin Peden, Eriko Kage-Nakadai, Shohei Mitani, and Ding Xue. 2010. “Caenorhabditis Elegans Transthyretin-like Protein TTR-52 Mediates Recognition of Apoptotic Cells by the CED-1 Phagocyte Receptor.” *Nature cell biology* 12: 655–64.

Williamson, Angela L, Sara Lustigman, Yelena Oksov, Vehid Deumic, Jordan Plieskatt, Susana Mendez, Bin Zhan, Maria Elena Bottazzi, Peter J Hotez, Alex Loukas, and I Nfect I Mmun. 2006. “Ancylostoma Caninum MTP-1 , an Astacin-Like Metalloprotease Secreted by Infective Hookworm Larvae , Is Involved in Tissue Migration.” 74(2): 961–67.

Yan, Yutao, Shuxian Liu, Guangcheng Song, Yixin Xu, and Colette Dissous. 2005. “Characterization of a Novel Vaccine Candidate and Serine Proteinase Inhibitor from Schistosoma Japonicum (Sj Serpin).” *Veterinary Parasitology* 131: 53–60.

Zarowiecki, Magdalena, and Matt Berriman. 2014. “What Helminth Genomes Have Taught Us about Parasite Evolution.” *Parasitology*: 1–13. http://www.ncbi.nlm.nih.gov/pubmed/25482650 (January 22, 2015).

Zheng, Yadong. 2013. “Phylogenetic Analysis of the Argonaute Protein Family in Platyhelminths.” *Molecular phylogenetics and evolution* 66(3): 1050–54. http://www.ncbi.nlm.nih.gov/pubmed/23211720 (April 13, 2015).